STA 380, Part 2: Exercises

Andrew Gillock, Ben Kanarick, Pranav Cheruku, Denise Neuman

2022-08-15

LINK TO GITHUB REPO: https://github.com/andrewgillock/groupExercises

Question 1: Probability Practice

Part A

RULE OF TOTAL PROBABILITY

$$P(Y) = P(TC) * P(Y|TC) + P(RC) * P(Y|RC)$$

$$0.65 = 0.7 * P(Y|TC) + 0.3 * 0.5$$

$$P(Y|TC) = \frac{(0.65 - (0.3 * 0.5))}{0.7}$$

The fraction of people who are truthful clickers and answered yes is 0.714.

Part B

BAYES' RULE

$$P(YesD|Pos) = (P(Pos|YesD) * \frac{P(YesD))}{P(Pos)}$$

RULE OF TOTAL PROBABILITY

$$P(Pos) = P(YesD)\text{ů}P(Pos|YesD) + P(NoD)\text{ů}P(Pos|NoD)$$

$$P(Pos) = (0.000025\text{ů}0.993) + ((1 - 0.000025)\text{ů}(1 - 0.9999))$$

$$probPos1_B = (0.000025 * 0.993) + ((1 - 0.000025) * (1 - 0.9999))$$

$$P(YesD|Pos) = \frac{(0.993 * 0.000025)}{probPos1_B}$$

The probability that someone has the disease given that they tested positive is 0.199.

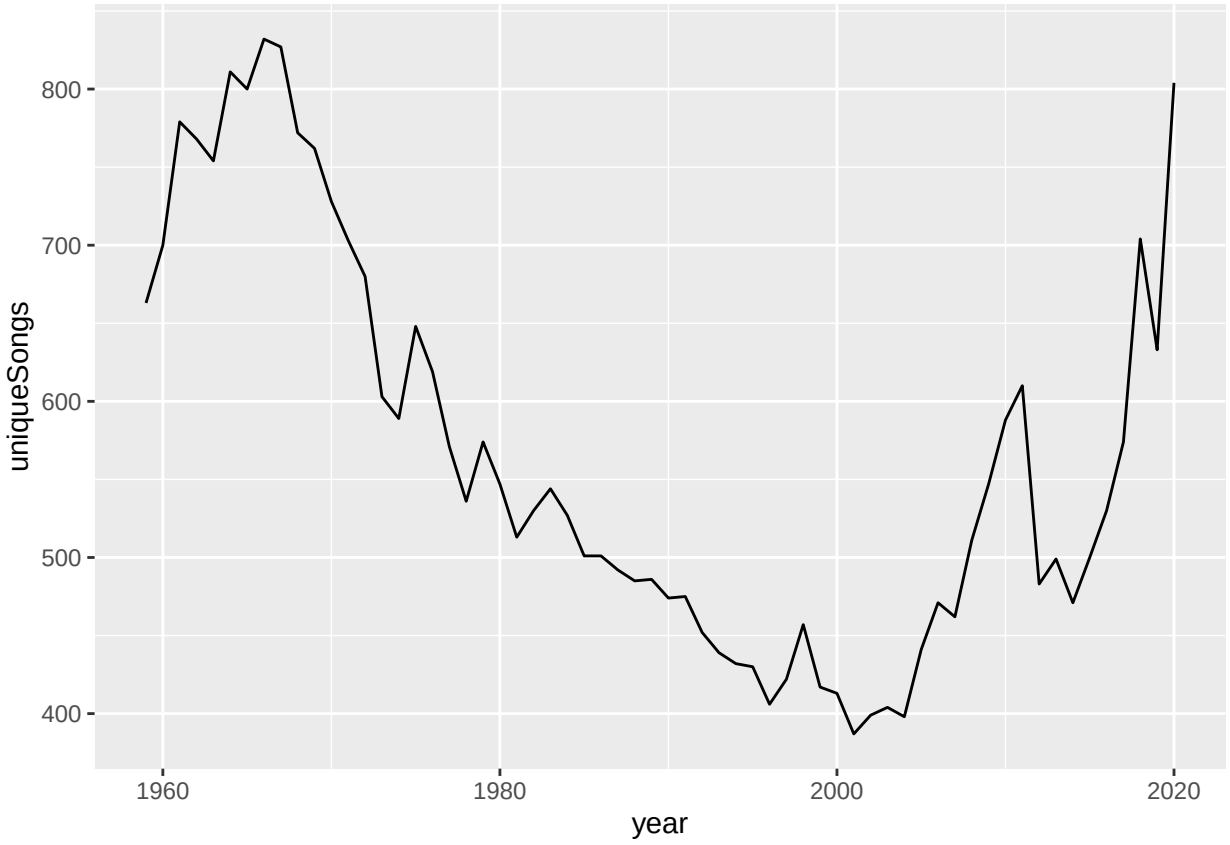
Question 2: Wrangling the Billboard Top 100

Part A

```
## # A tibble: 10 x 3
## # Groups:   performer [10]
##    performer                                song                          count
##    <chr>                                    <chr>                         <int>
##  1 Imagine Dragons                          Radioactive                      87
##  2 AWOLNATION                               Sail                             79
##  3 Jason Mraz                               I'm Yours                        76
##  4 The Weeknd                               Blinding Lights                  76
##  5 LeAnn Rimes                              How Do I Live                    69
##  6 LMFAO Featuring Lauren Bennett & GoonRock Party Rock Anthem               68
##  7 OneRepublic                              Counting Stars                   68
##  8 Adele                                    Rolling In The Deep              65
##  9 Jewel                                    Foolish Games/You Were Meant~    65
## 10 Carrie Underwood                         Before He Cheats                 64
```

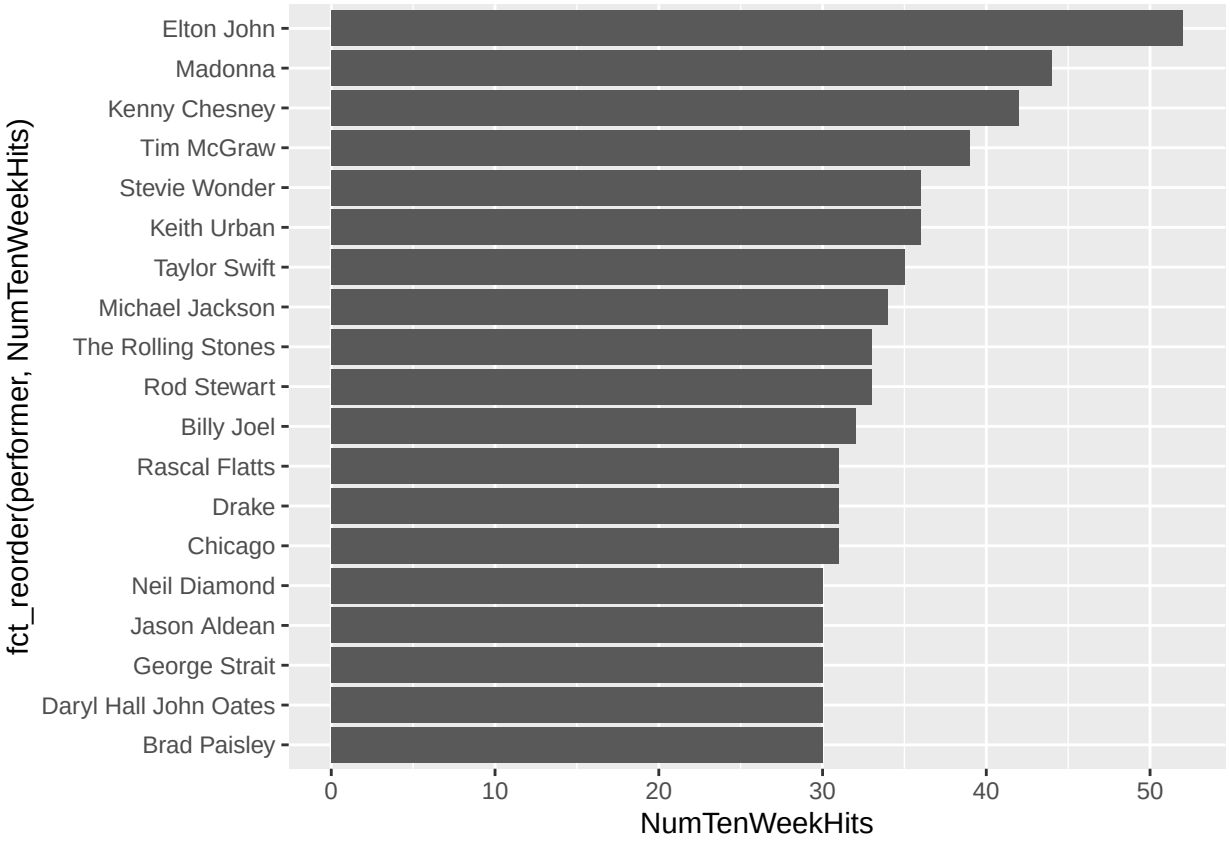
The table shows the top 10 songs since 1958 as measured by the total number of weeks that a song spent on the Billboard Top 100. The data shows that Radioactive by Imagine Dragons has been 87 weeks in the Billboard top 100, making it the song that has been more weeks in this ranking since 1958 (and up to week 22 of 2021).

Part B



In the plot, it can be seen that the musical diversity of the Billboard top 100 declined approximately 58.82% (850 to 380 unique songs) from the mid 60s to the early 2000s. After that, the diversity increased again by approximately 110.53% (380 to unique 800 songs) in the last 20 years, having a downfall in the early 2010s and re-surging the following years. The plot shows that in the year 2020, the musical diversity of the Billboard Top 100 was only 5.88% lower (850 vs 800 unique songs) than in the mid 60s.

Part C



After getting the songs that were in the top 100 for more than 10 weeks and getting the performers with 30+ Ten-Week Hits, Fig Q2C was created. The plot shows that Elton John is the performer with the most Ten Week Hits in the Billboard Top 100 since 1958, with 52 unique songs. This is 18.18% higher (44 vs 52 unique songs) than Madonna which is the second artists with more Ten Week Hits.

Question 3: Visual Story Telling Part 1: Green Buildings

So, on the face of it, the median profit per square feet for a Green building is 24.4337, which is higher than the median profit per square feet for a normal building is 21.3641, which means that the profit per square feet for a green building is higher than the profit per square feet for a normal building. We use median to remove outliers that might skew the data.

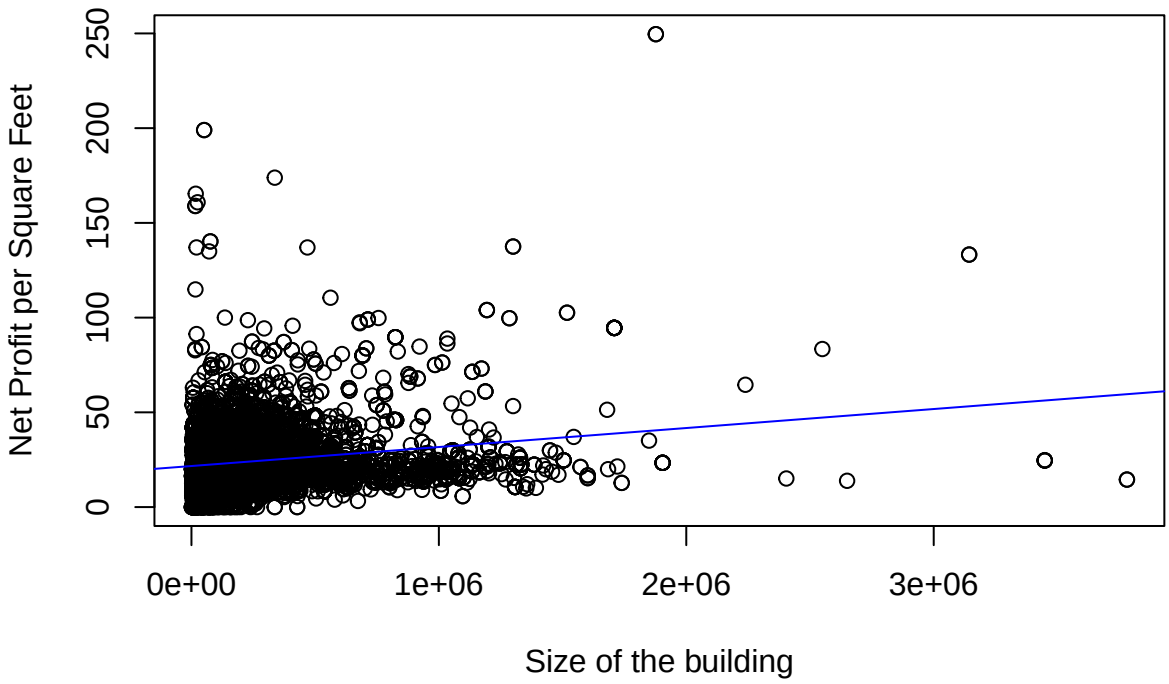
Now, we can see whether the normal or green buildings will become more profitable. We have the median profit per sqft for green and normal, we have the minimum number of years they are expected to earn rent, and the we have the total cost of construction, and converting a building to a green one. So, we can calculate the net profit.

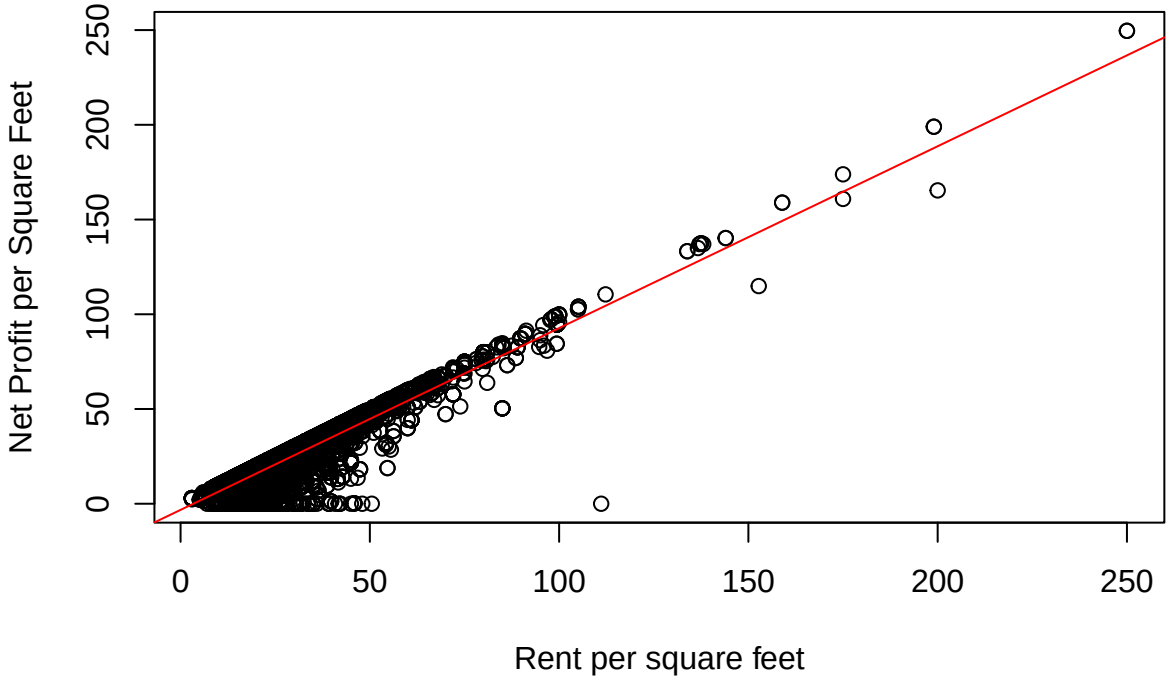
We see that the net profit for the green building is 78252525 over a 30-yr period. This is higher than the net profit for the normal building over a 30-yr period of 60232325. Therefore, it seems like a good investment decision to proceed with.

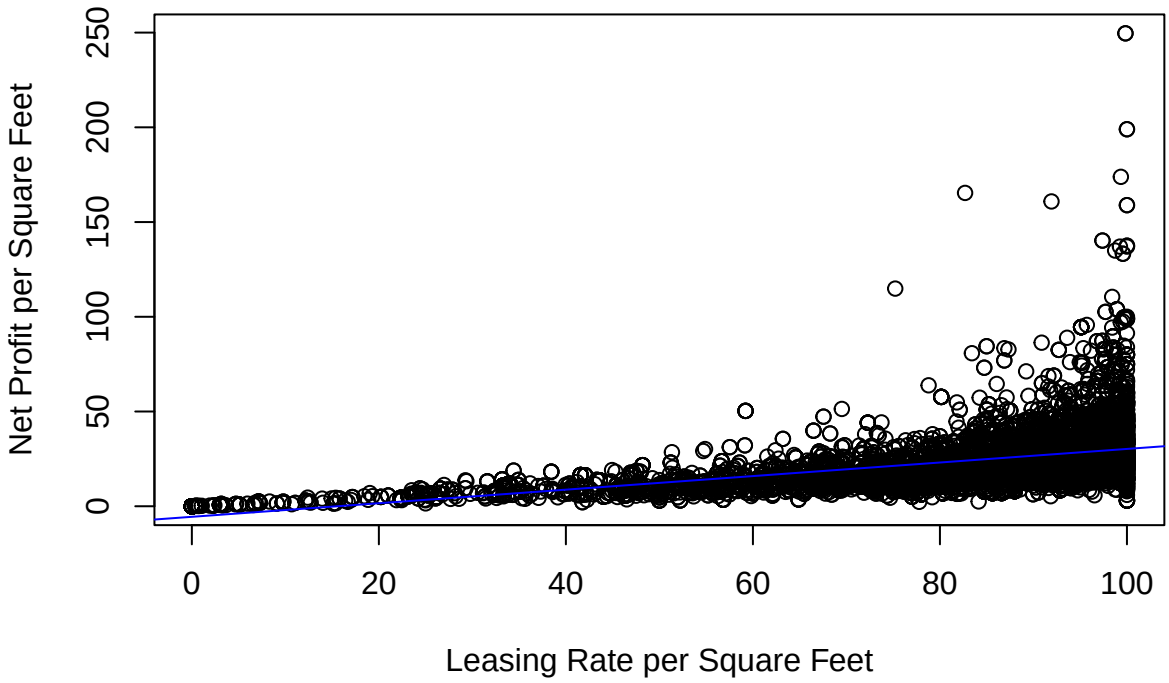
But why is this? Could it really be that green buildings are more desirable and profitable, or could there be others factors cofounding the situation. Looking at the analysis, he takes into consideration one variable, low occupancy rates, and then looks at the median rent. We will need to adjust for factors.

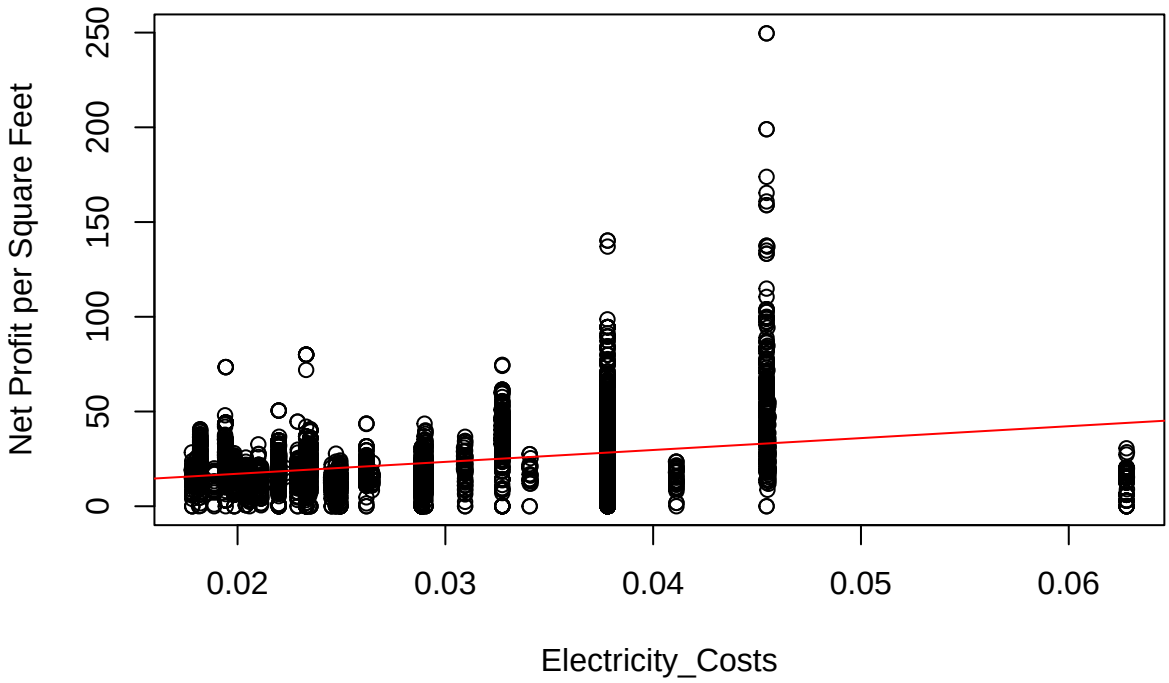
To illustrate the impact of how factors can skew the relationship between the building type(green or normal) on the profitability, we will graph the impact of some predictors on profitability. That is, the more of a predictor, does profitability go up or down?

Including Plots

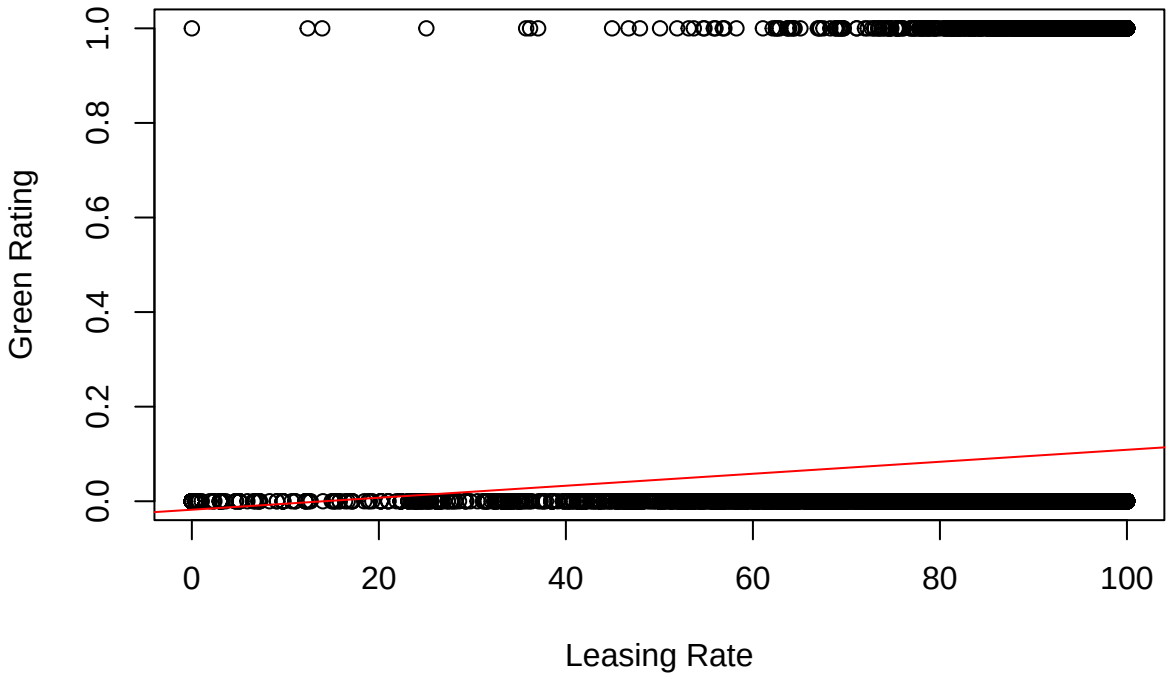





Net Profit per Square Feet
0
50
100
150
200
250
0
20
40
60
80
100
Leasing Rate per Square Feet



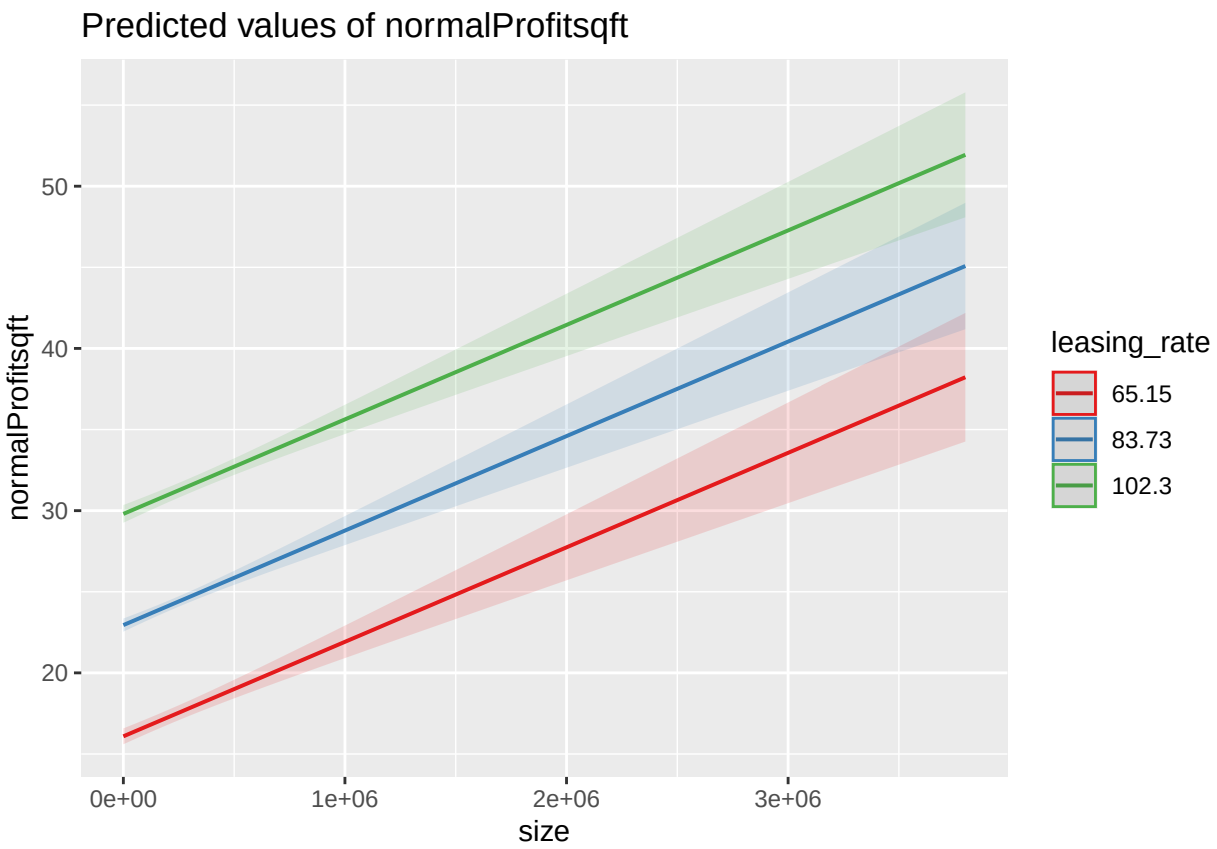
So, we can see that the graphed predictors have some impact on the Net Profit per Square Feet, so we need to take the impact of those predictors into account to find out whether a green building is more profitable than a normal building. For instance, in the analysis provided, the individual removed the bottom 10% of occupancy rates. That skewed the data against Green Buildings, as we can see in the plot below:



Areas with higher leasing rates were more likely to have a Green Rating than buildings with lower leasing rates. Removing lower buildings with occupancy rates below 10% skewed the data against Green Ratings, by removing some low performing buildings which were disproportionately normal and not Green. This is an example of the impact that predictors can have on the stated performance of Green vs Normal. We can adjust by doing our analysis on a narrow slice of buildings with very similar occupancy rates, rather than excluding a small subsection and keeping the rest. This can also be done for all of the different variables.

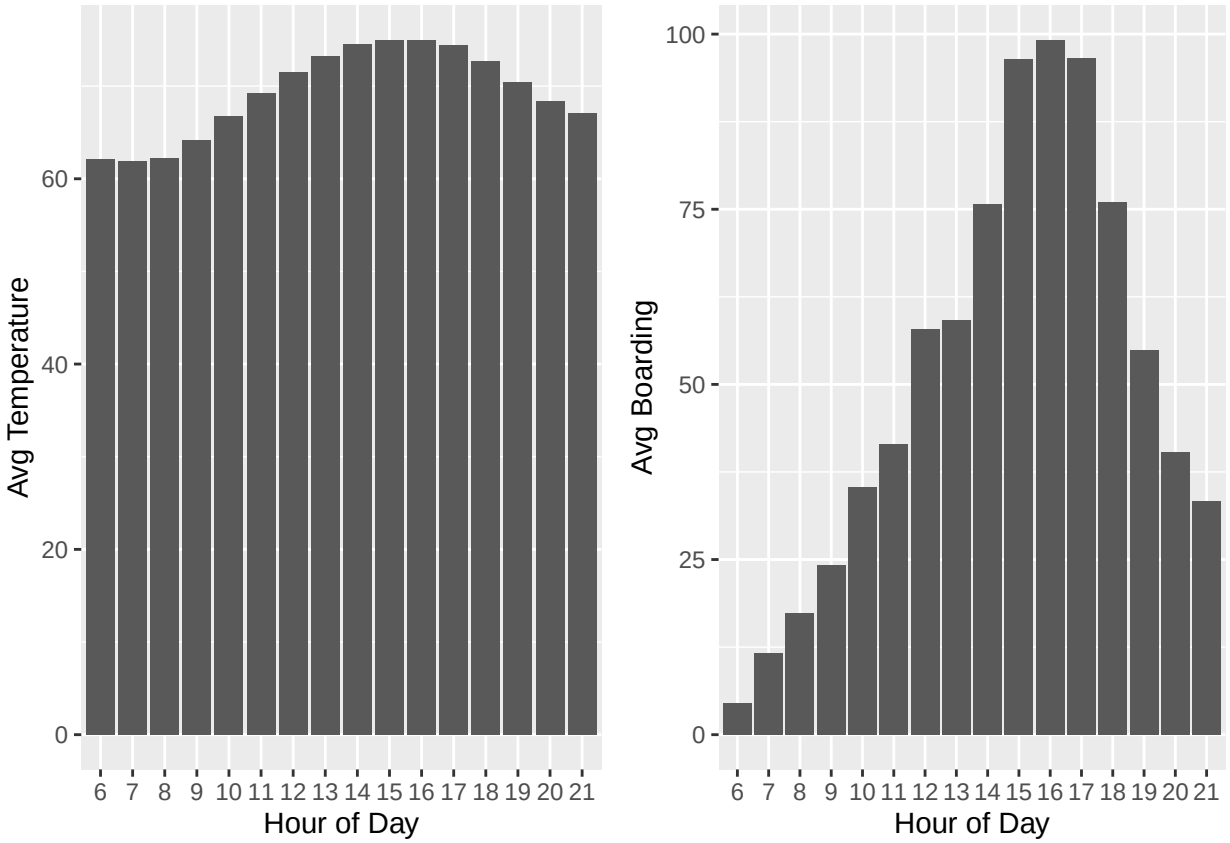
For instance, we are going to adjust for size and leasing rate, and compare net profit per square ft for green and normal buildings. So, if a building has a similar size and occupancy rate, will it have a higher profit per sqft if it was green rather than normal?


Predicted values of greenProfitsqft
greenProfitsqft
35
30
25
20
15
0
500000
1000000
1500000
size
leasing_rate
77.59
89.41
101.24

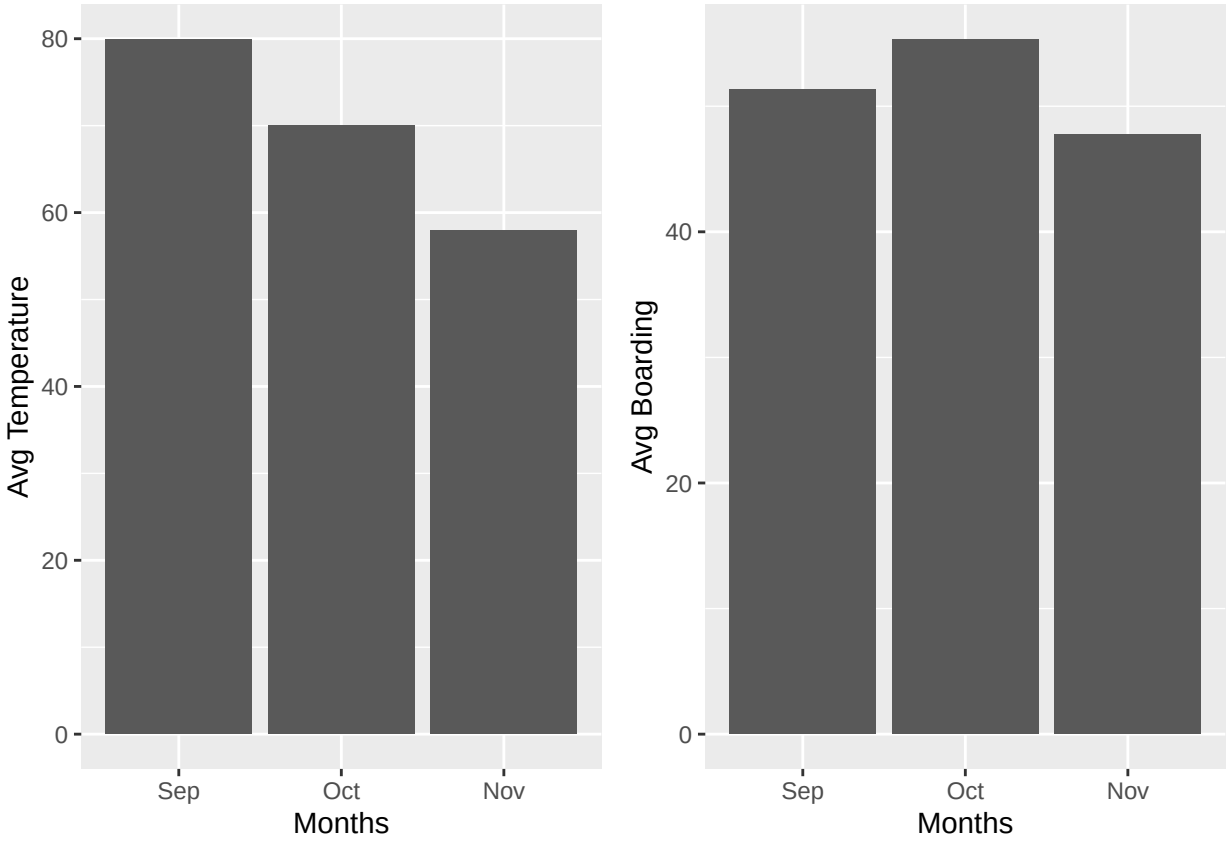


Shockingly, we can see that for similar size and leasing rates, normal buildings actually have higher profit per sqft than green buildings. That means that perhaps there are cofounding variables that impact net profit per sqft that mask the fact that normal buildings are actually more profitable than green buildings. Or it could be that there are other variables that we are not taking into account, that if accounted for, could show green buildings as more profitable overall. Regardless, it is clear that the analysis that the analyst did is insufficient.

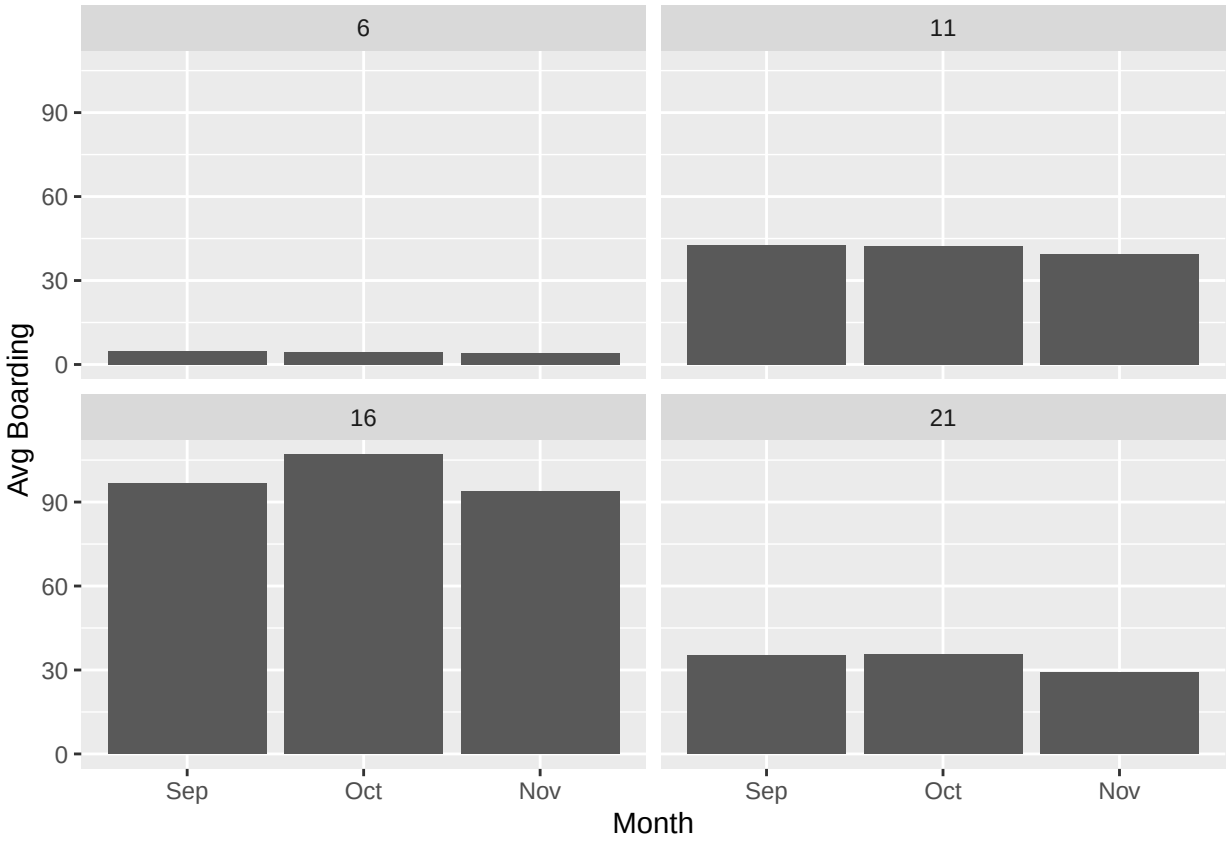
Question 4: Visual Story Telling Part 2: Capital Metro



Bar plot of average boarding through the hour of day. The plot shows that more people board the Capital Metro transport system when the temperature is higher



Bar plot of average boarding through the months. The plot shows that even though the average temperature lowers from Sep to Oct and from Oct to Nov, the average number of people boarding the public transport system increases 11.4% (98,530 to 109,767 boardings) from Sep to Oct and lowers 16.45% (109,767 to 91,707) from Oct to Nov.



Average boarding through the hours of day in different months. Tha graphs show that in the afternoon more people board the Capital Metro transport system and that early morning is when less people use the transport system.

In a day to day basis, people prefer to board the Capitol Metro Transport System when the temperature is higher. However, in the bigger picture, as the temperature lowers, more people use the transport system. November, seems to have lower number of boardings, but this may be so because of Thanksgiving break, when students go away for a week.

Question 5: Portfolio Modeling

Portfolio 1:

We have decided to use 5 very different ETFs for this problem, they are listed and imported below: This portfolio consists of primarily safe and risk-averse ETFS.

SPY: SPDR S&P 500 ETF Trust

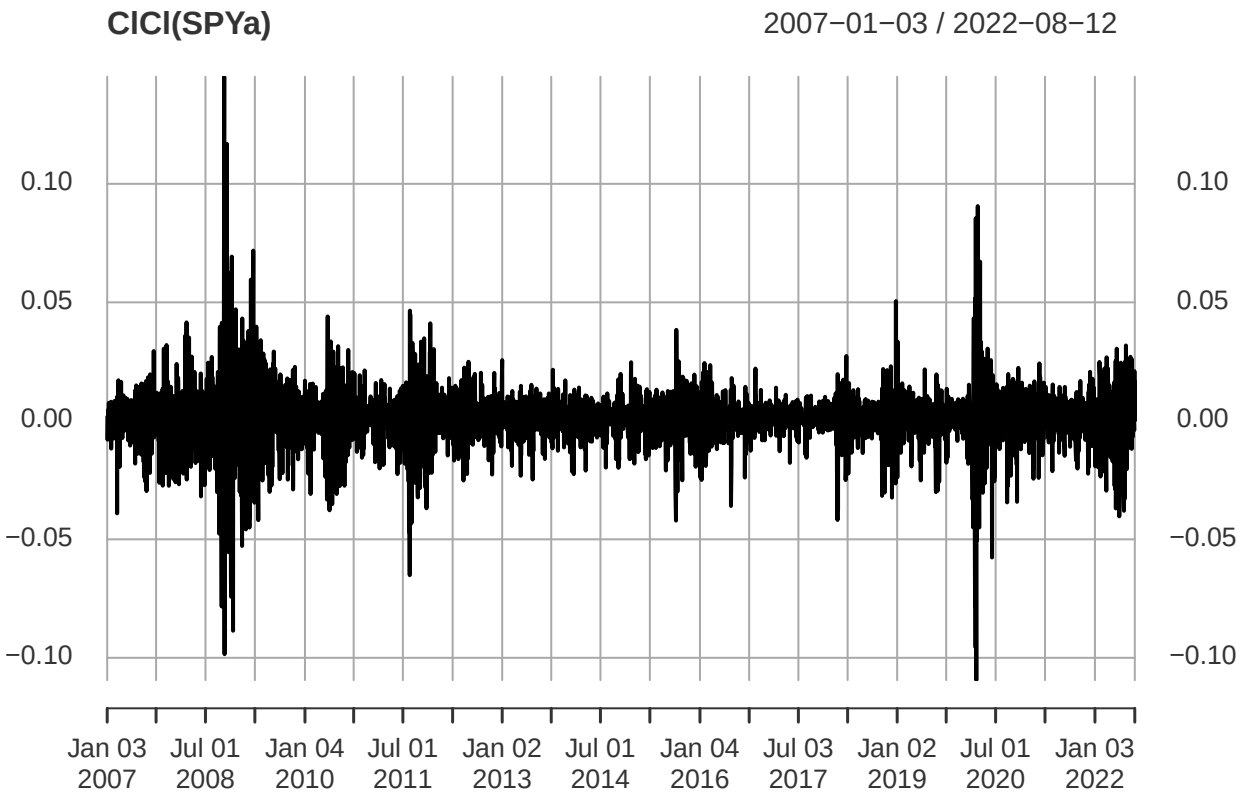
HYG: iShares iBoxx $ High Yield Corporate Bond ETF

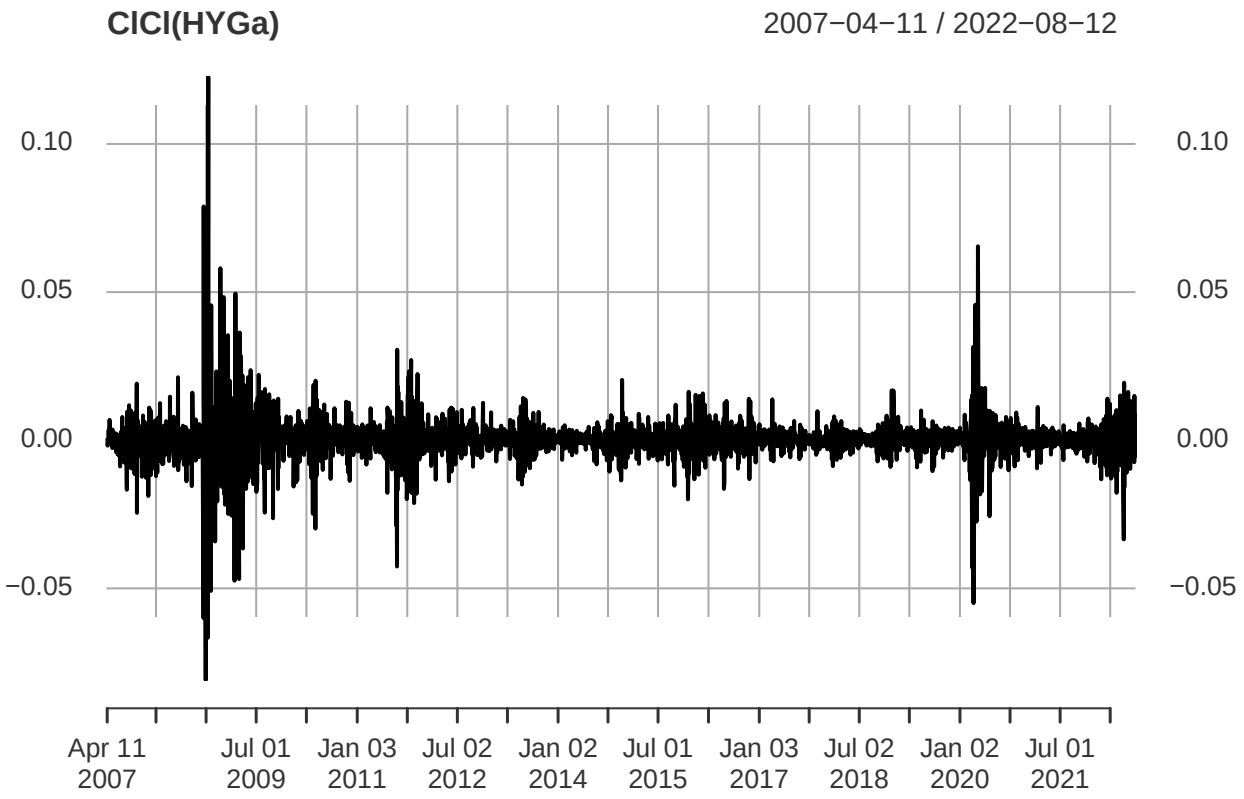
DBA: Invesco DB Agriculture Fund

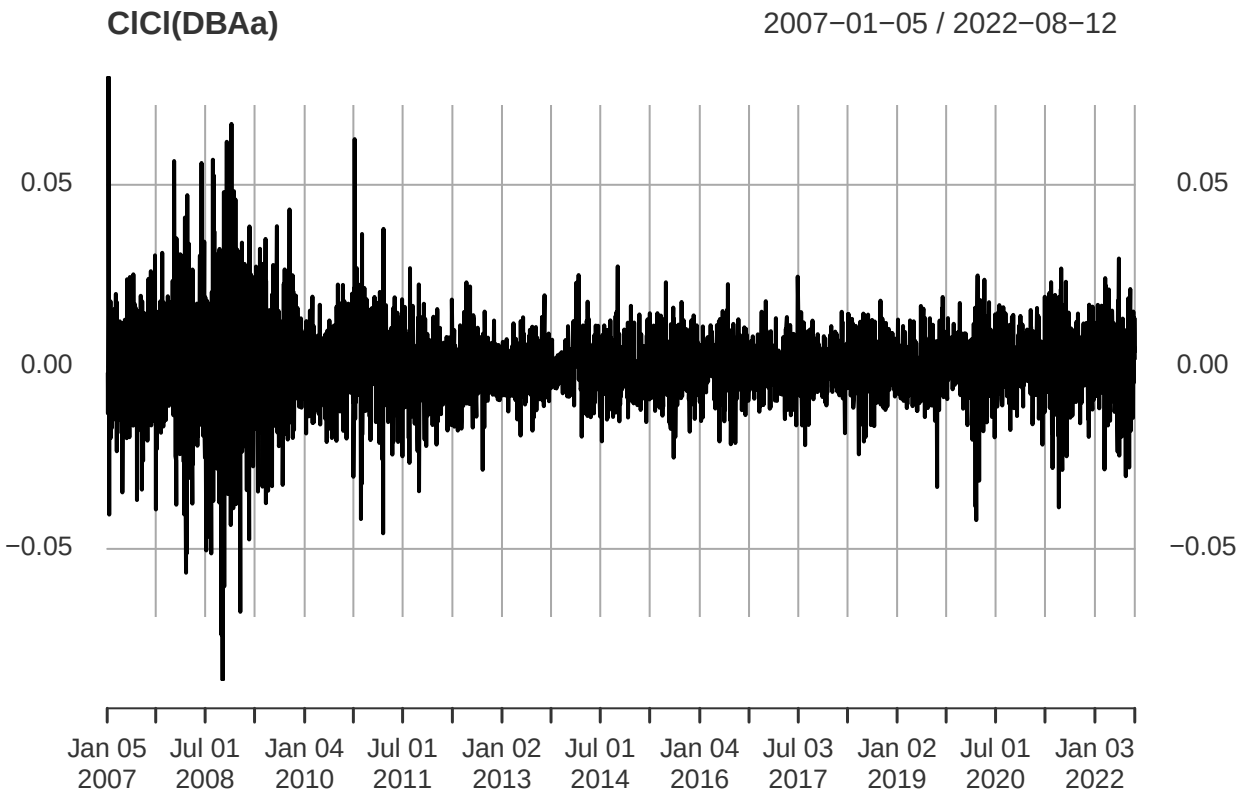
UVXY: ProShares Ultra VIX Short-Term Futures ETF

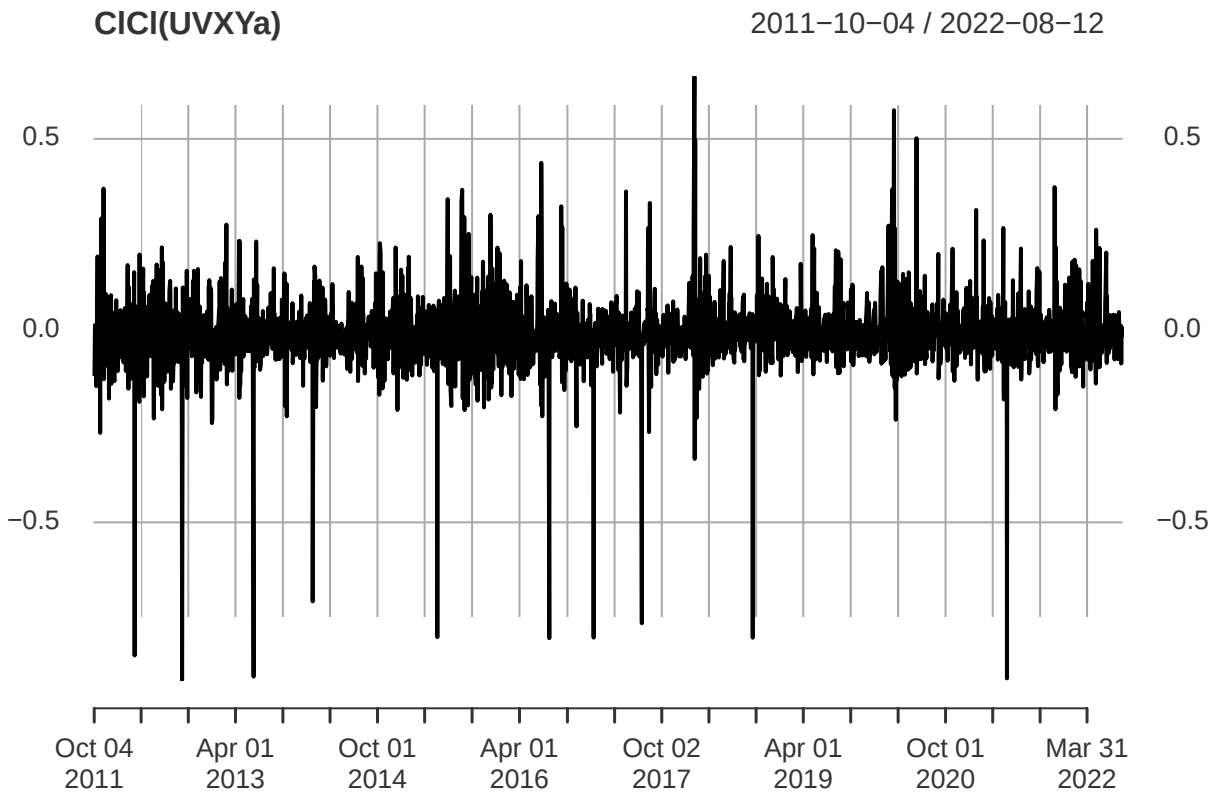
FXE: Invesco CurrencyShares Euro Trust

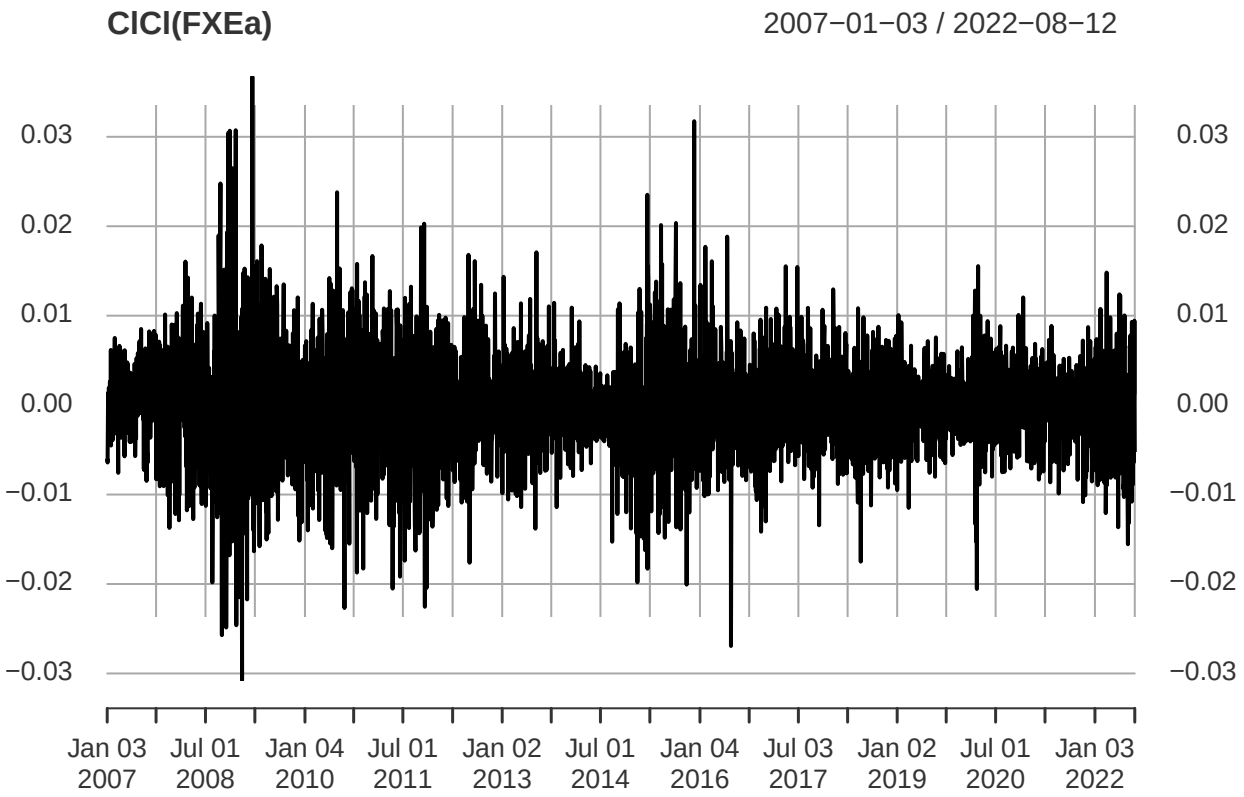
Adjust for splits and dividends Now that the data has been adjusted for splits and dividends, the following graphs display the close-to-close changes for each of the ETFs in our portfolio.


ClCl(SPYa)
2007−01−03 / 2022−08−12
0.10
0.05
0.00
−0.05
−0.10
Jan 03 2007
Jul 01 2008
Jan 04 2010
Jul 01 2011
Jan 02 2013
Jul 01 2014
Jan 04 2016
Jul 03 2017
Jan 02 2019
Jul 01 2020
Jan 03 2022

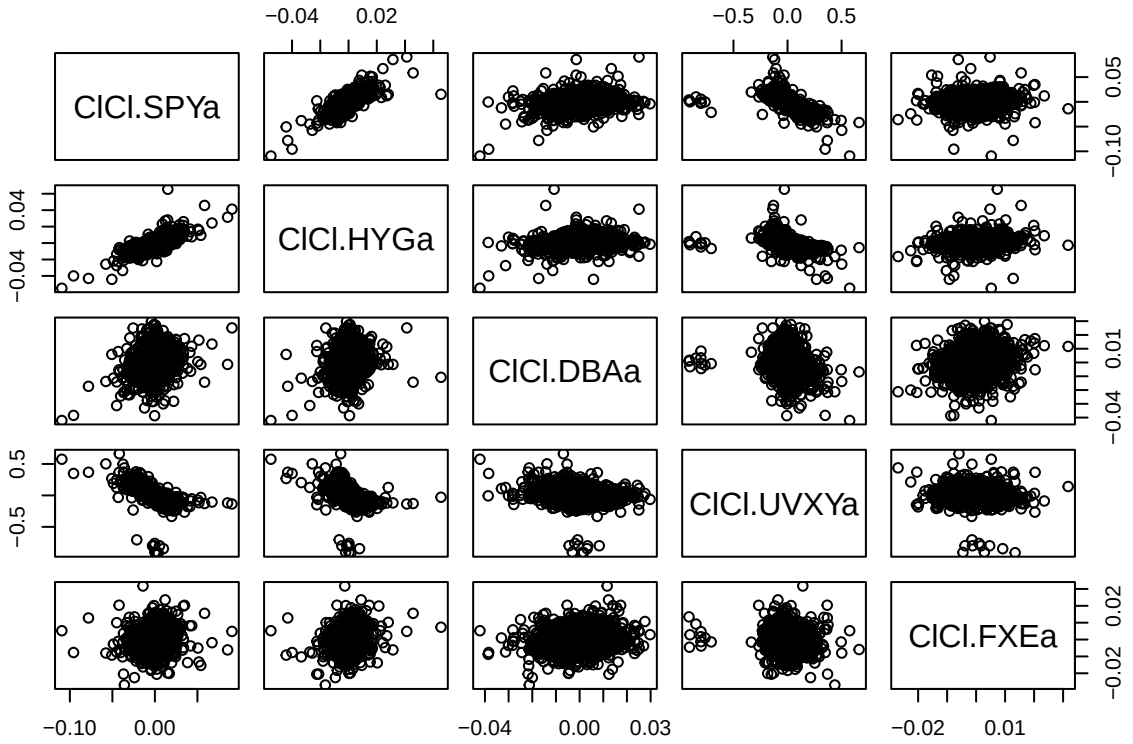

ClCl(HYGa)
2007−04−11 / 2022−08−12
0.10
0.05
0.00
−0.05
Apr 11 2007
Jul 01 2009
Jan 03 2011
Jul 02 2012
Jan 02 2014
Jul 01 2015
Jan 03 2017
Jul 02 2018
Jan 02 2020
Jul 01 2021


ClCl(DBAa)
2007−01−05 / 2022−08−12
0.05
0.00
−0.05
Jan 05 2007
Jul 01 2008
Jan 04 2010
Jul 01 2011
Jan 02 2013
Jul 01 2014
Jan 04 2016
Jul 03 2017
Jan 02 2019
Jul 01 2020
Jan 03 2022


ClCl(UVXYa)
2011−10−04 / 2022−08−12
0.5
0.0
−0.5
Oct 04
2011
Apr 01
2013
Oct 01
2014
Apr 01
2016
Oct 02
2017
Apr 01
2019
Oct 01
2020
Mar 31
2022

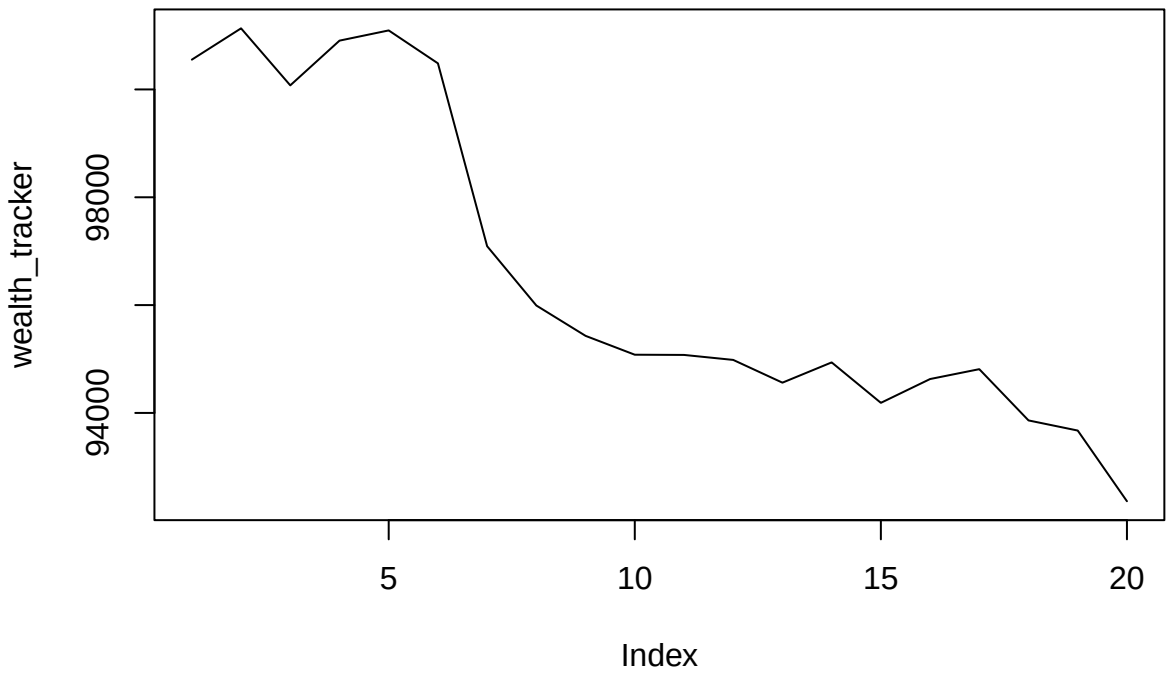


Combine close to close changes in a single matrix Omit all NA in each set so all ETFS start at the same date Once we have found a common start date for the portfolio and got rid of NA values, we are now able to use the pairs function to show the correlation between each of the ETFs in our portfolio.

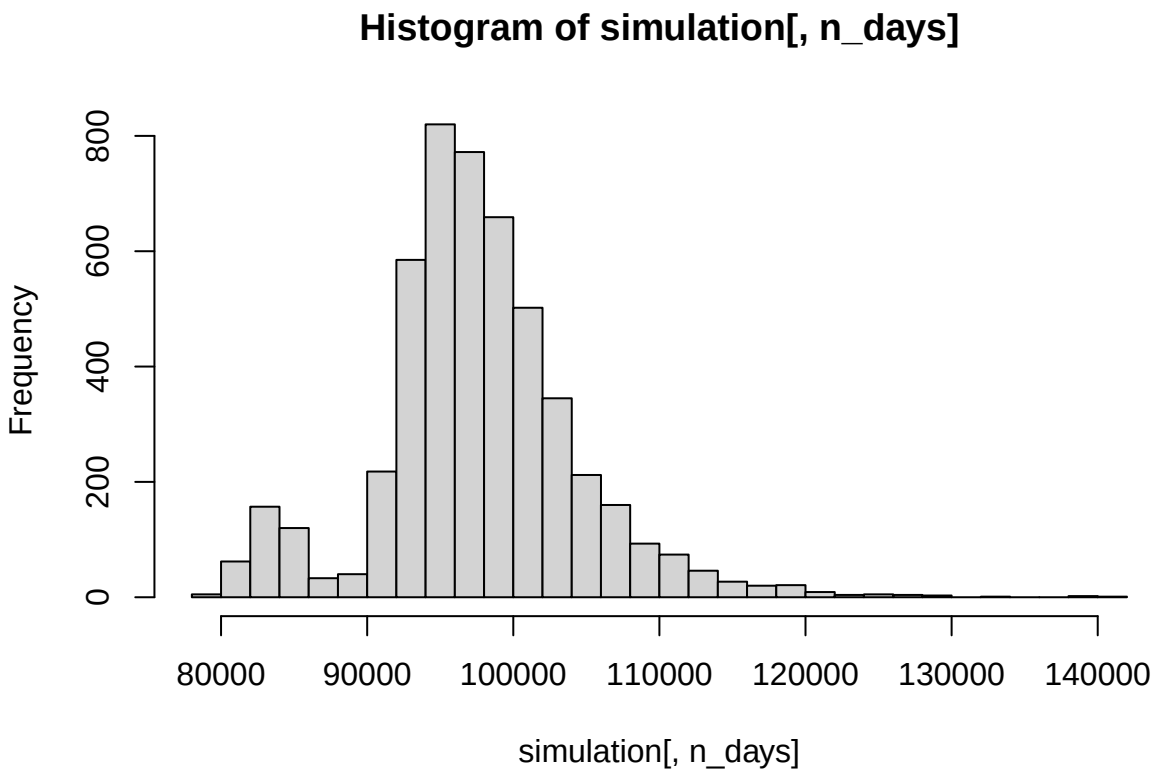


Begin bootstrap approach:

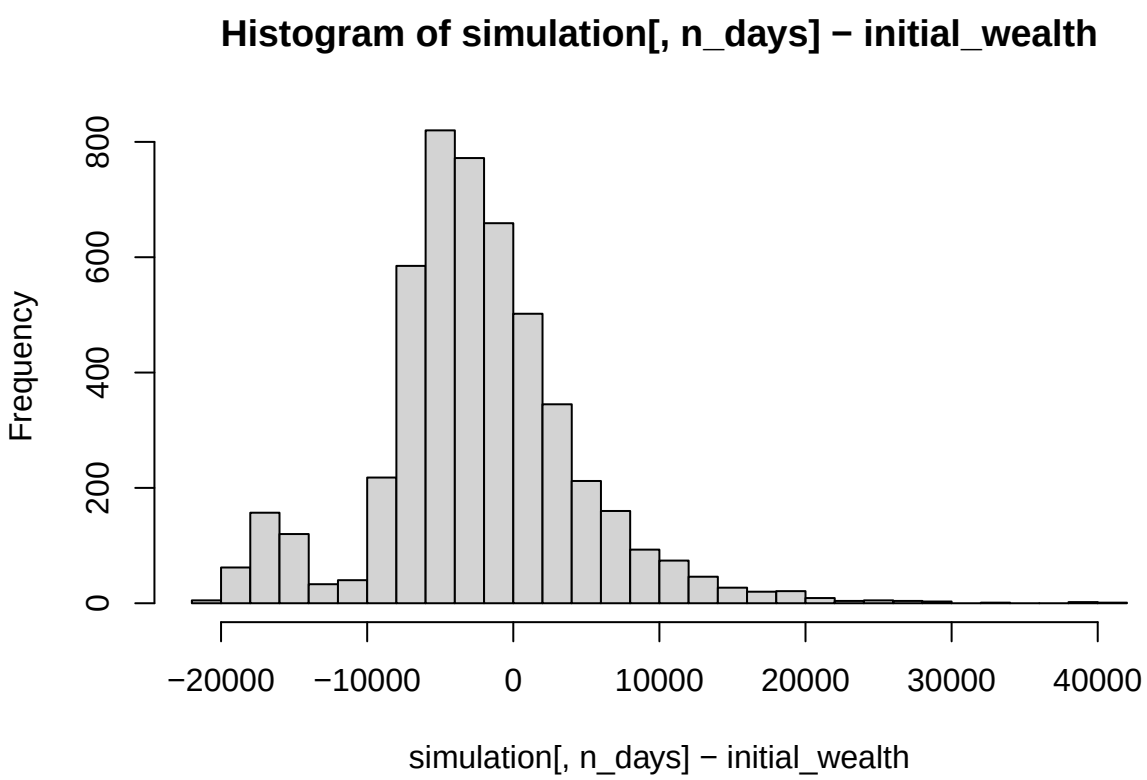
Sample random return from the empirical joint distribution This simulates a random day Update the value of the holdings to be equal weighted Since we chose 5 ETFS each one will have a weight of 20% Compute our new total wealth Now we will loop over 4 trading weeks, or 20 trading days using the same logic as mapped out above. The graph below shows the average of the 5000 trading simulations over the 20 day time frame as indicated by the question. We can see that this portfolio had an average value of 92010.53



Below is a distribution of each of the 5000 runs with respect to the final portfolio value after each run.



Below is a profit/loss distribution for each of the simulated 5000 portfolios over the 20 day trading span.



The 5% value at risk for this portfolio is 15756.17

Portfolio 1 Summary and Findings This portfolio consists of 5 equal-weighted ETFS as listed below:

SPY: SPDR S&P 500 ETF Trust

HYG: iShares iBoxx $ High Yield Corporate Bond ETF

DBA: Invesco DB Agriculture Fund

UVXY: ProShares Ultra VIX Short-Term Futures ETF

FXE: Invesco CurrencyShares Euro Trust

After analyzing our portfolio, we found the 5% VAR to be $15756.17. This value quantifies the extent of possible financial losses for our portfolio over the 20 trading day period simulation. For this particular instance of the simulation, the ending value of our portfolio was 92010.53 dollars.

Portfolio 2

We have decided to use 5 very aggressive ETFs for this problem, they are listed and imported below: This portfolio consists of risky and levered equity ETFs.

SOXL: Direxion Daily Semiconductor Bull 3x Shares

ROM: ProShares Utlra Technology

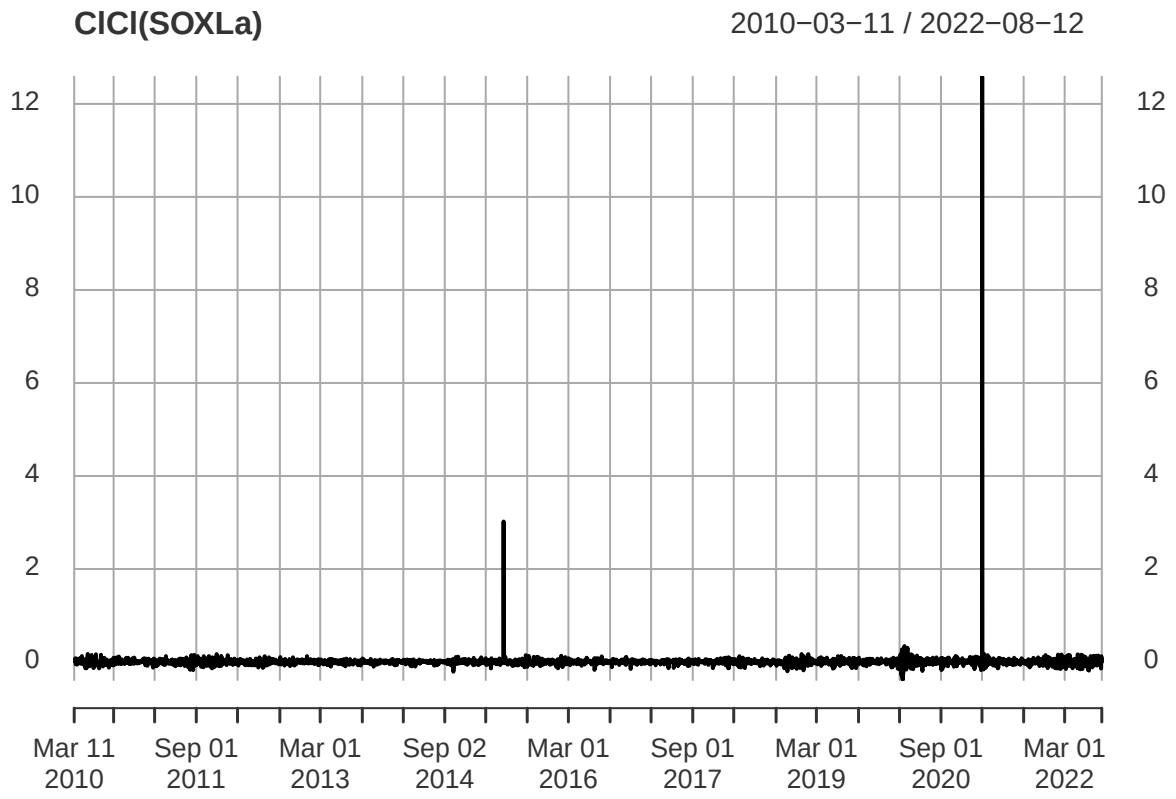
DPST: Direxion Daily Regional Banks Bull 3x Shares

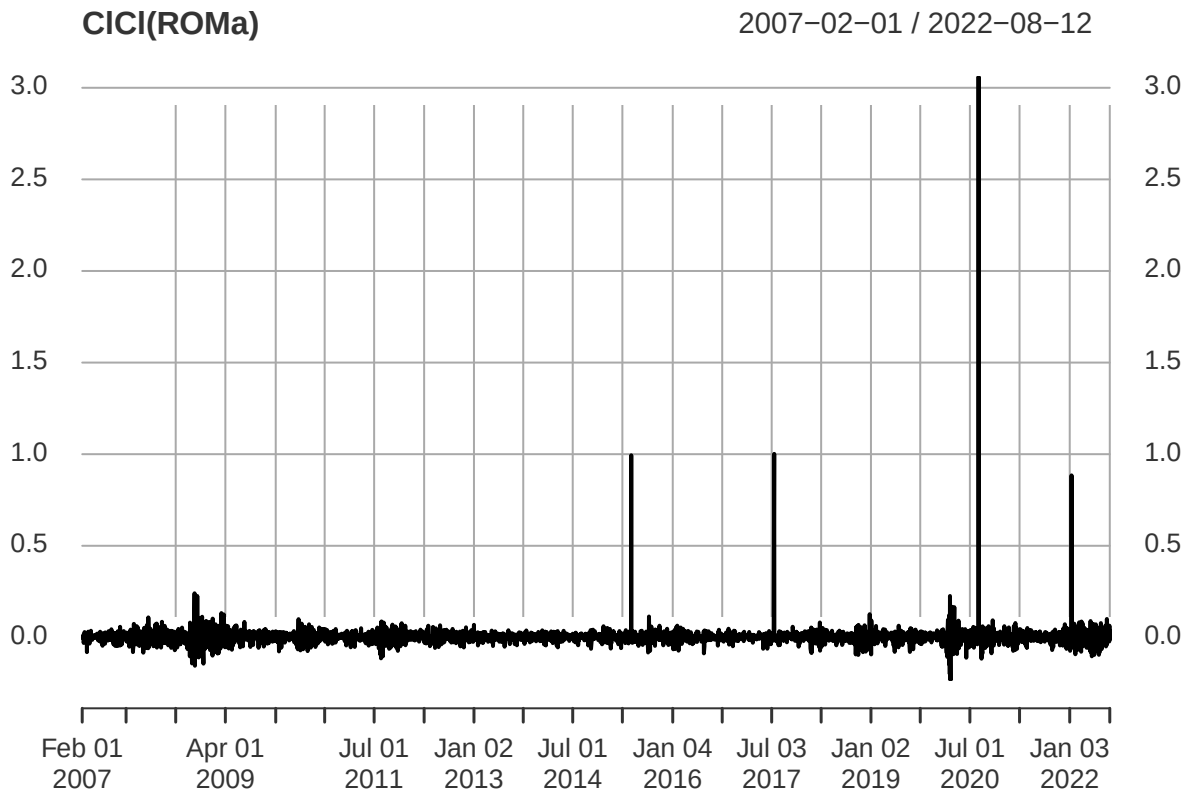
DRN: Direxion Daily Real Estate Bull 3x Shares

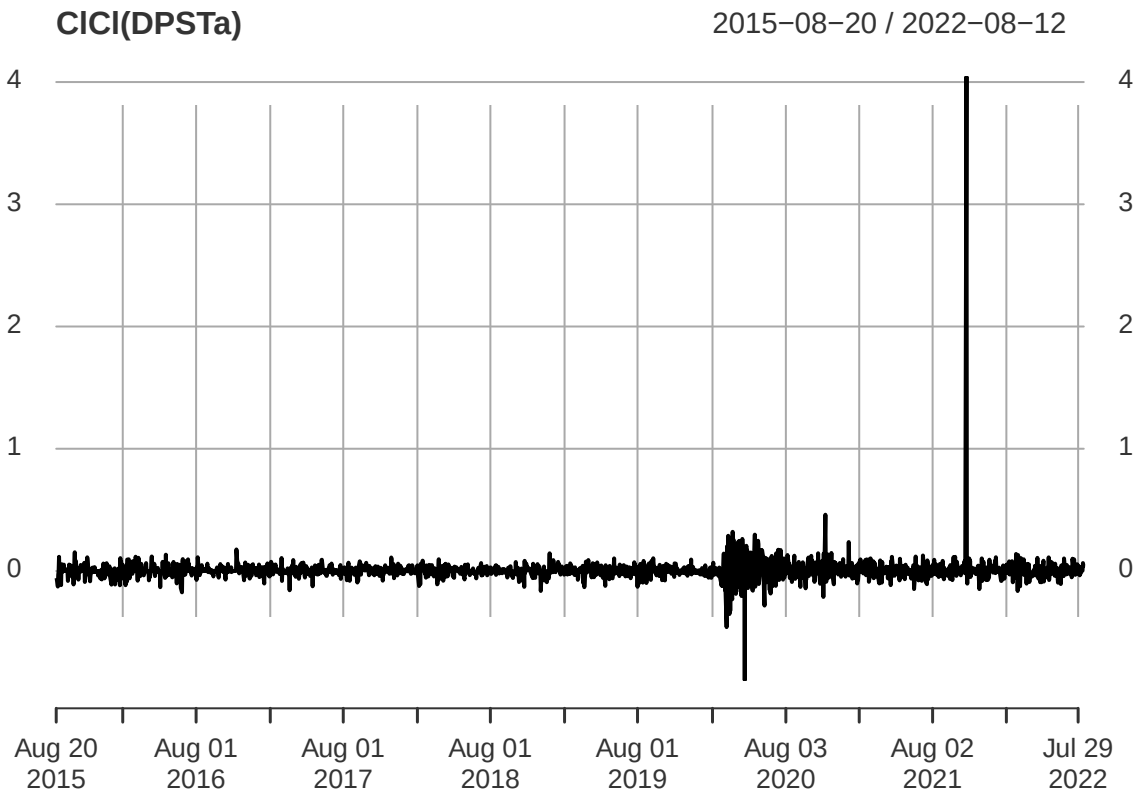
UCO: ProShares Ultra Bloomberg Crude Oil

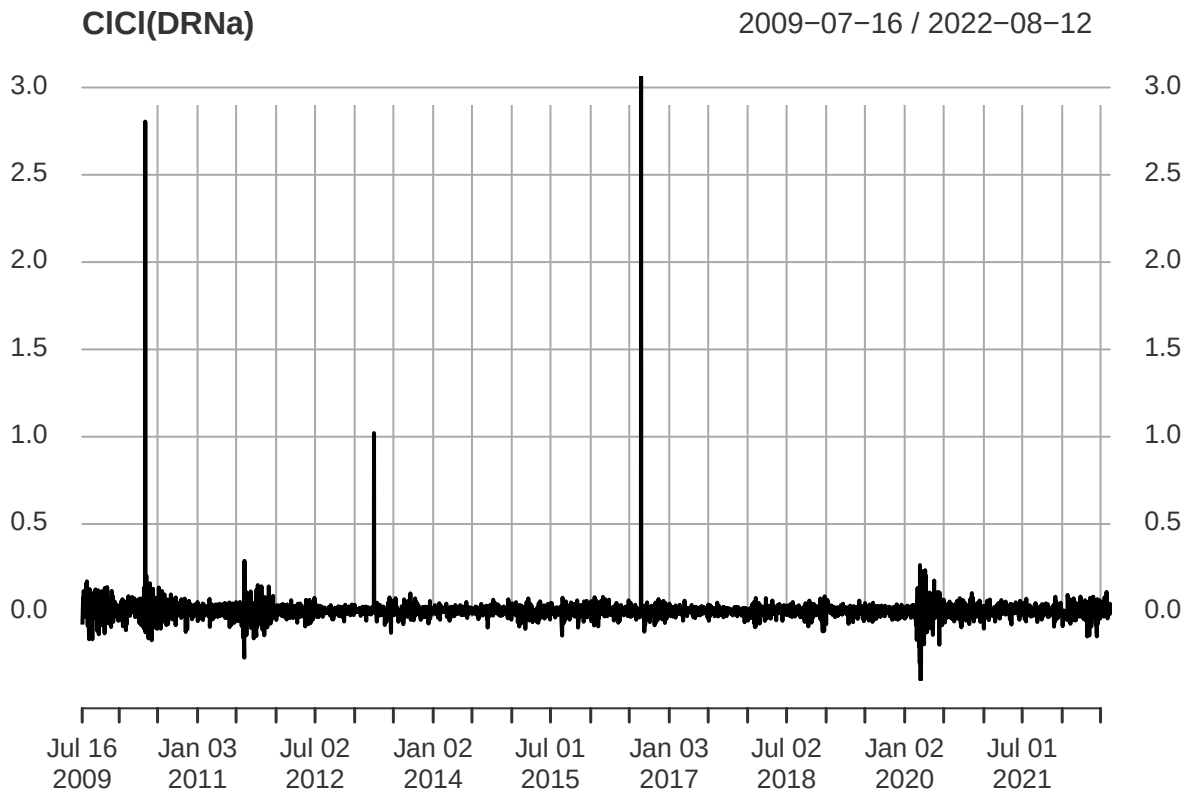
Adjust for splits and dividends

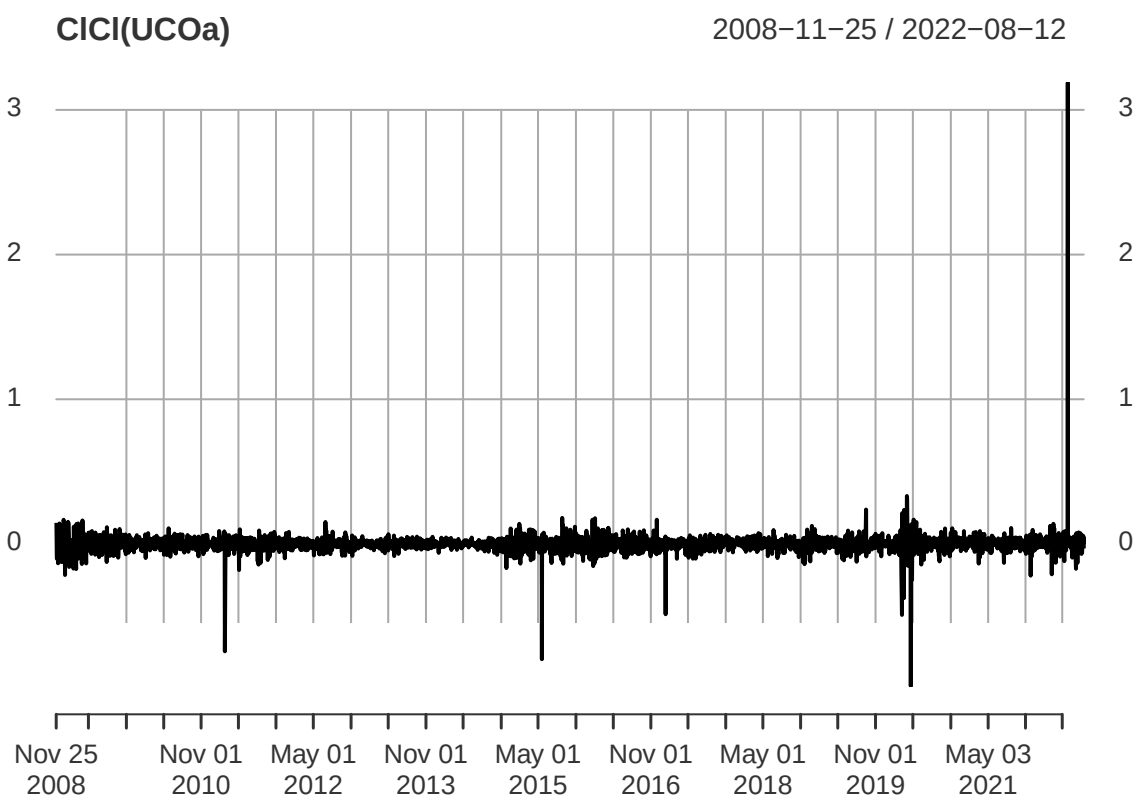
Now that the data has been adjusted for splits and dividends, the following graphs display the close-to-close changes for each of the ETFs in our portfolio.



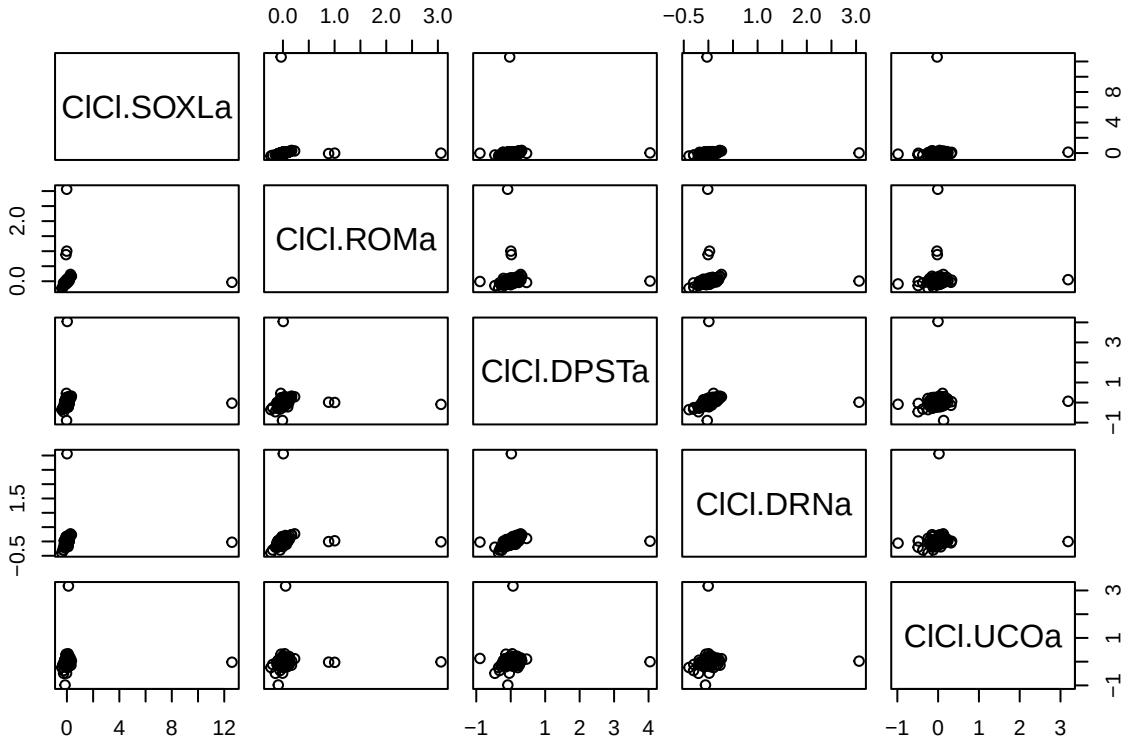

ClCl(ROMa)
2007−02−01 / 2022−08−12
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Feb 01 2007
Apr 01 2009
Jul 01 2011
Jan 02 2013
Jul 01 2014
Jan 04 2016
Jul 03 2017
Jan 02 2019
Jul 01 2020
Jan 03 2022


ClCl(DPSTa)
2015−08−20 / 2022−08−12
4
3
2
1
0
Aug 20
2015
Aug 01
2016
Aug 01
2017
Aug 01
2018
Aug 01
2019
Aug 03
2020
Aug 02
2021
Jul 29
2022


ClCl(DRNa)
2009−07−16 / 2022−08−12
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Jul 16 2009
Jan 03 2011
Jul 02 2012
Jan 02 2014
Jul 01 2015
Jan 03 2017
Jul 02 2018
Jan 02 2020
Jul 01 2021

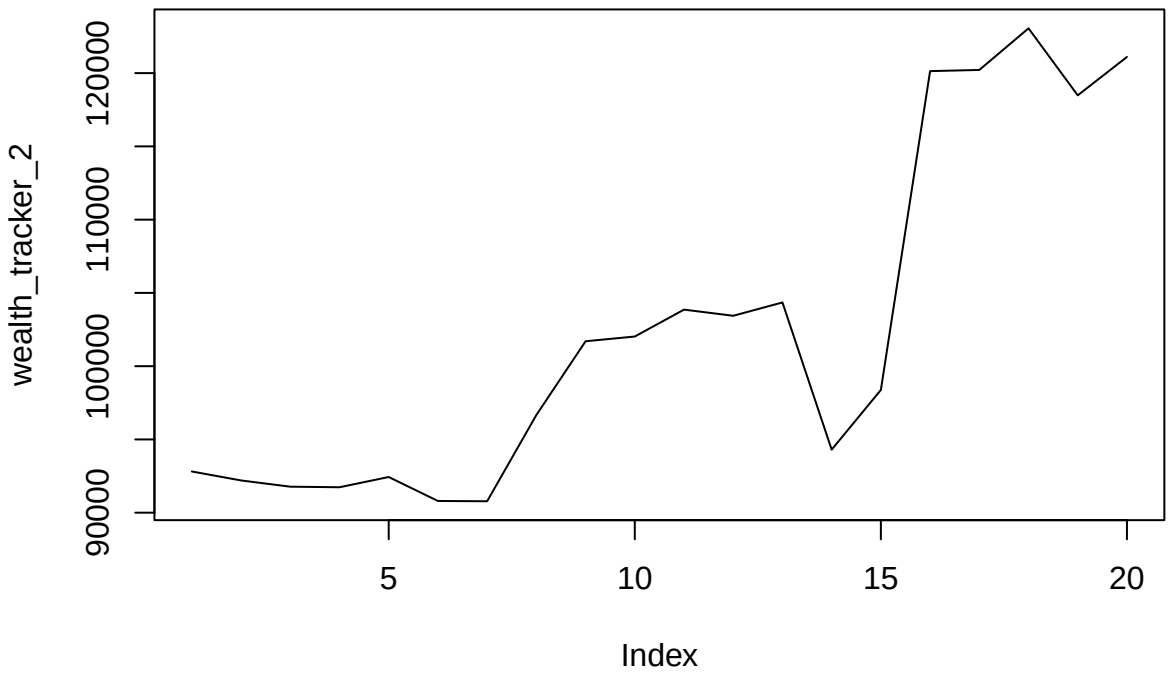


Combine close to close changes in a single matrix Omit all NA in each set so all ETFS start at the same date Once we have found a common start date for the portfolio and got rid of NA values, we are now able to use the pairs function to show the correlation between each of the ETFs in our portfolio.

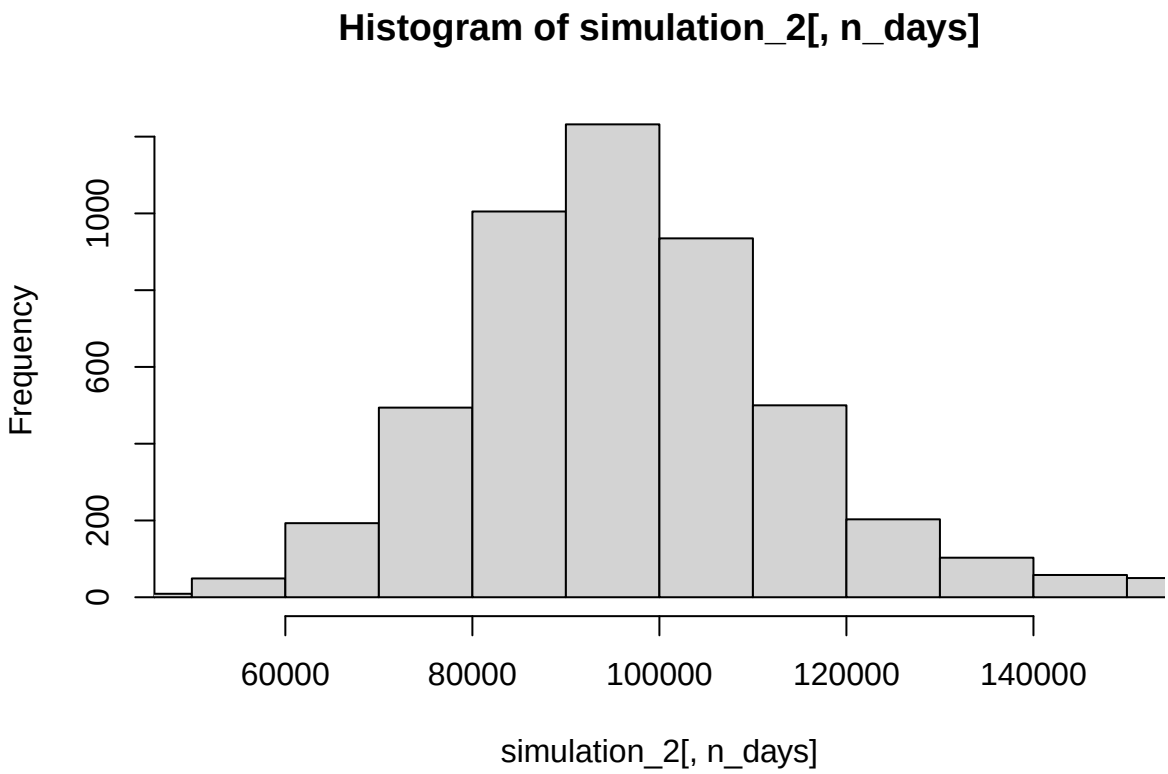


Begin bootstrap approach:

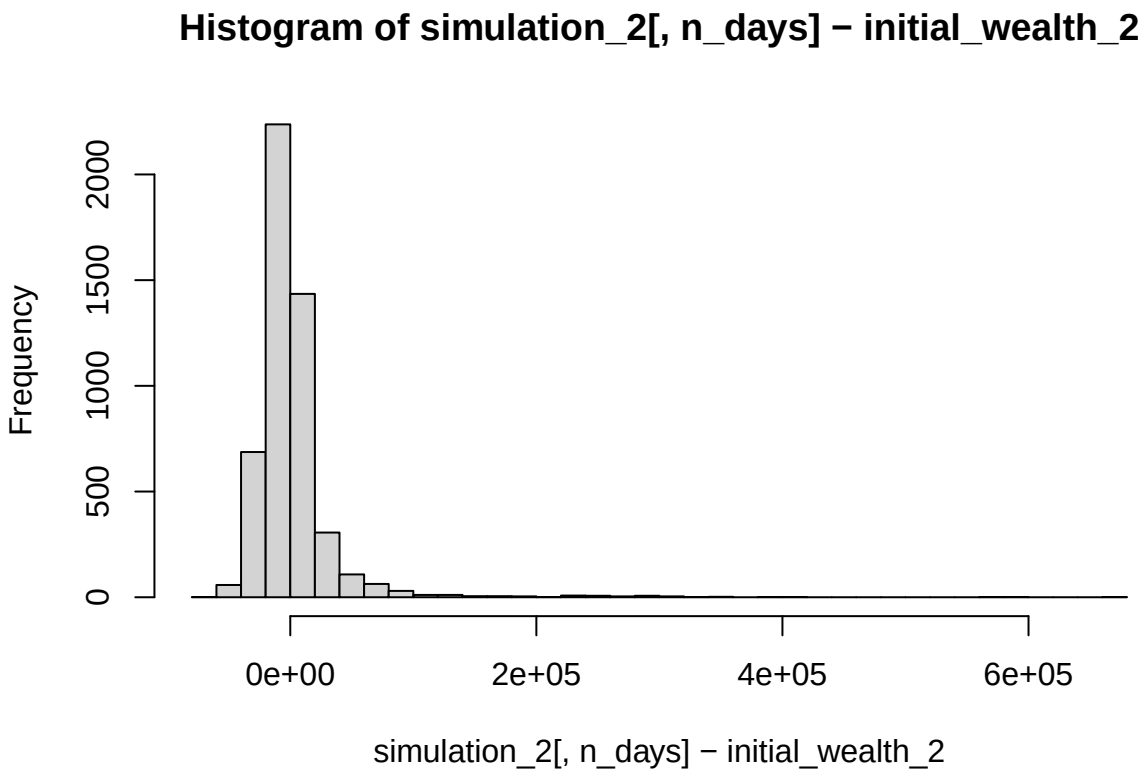
Sample a random return from the empirical joint distribution This simulates a random day Update the value of the holdings to be equal weighted Since we chose 5 ETFS each one will have a weight of 20% Compute our new total wealth Now we will loop over 4 trading weeks, or 20 trading days using the same logic as mapped out above. The graph below shows the average of the 5000 trading simulations over the 20 day time frame as indicated by the question. We can see that this portfolio had an average value of 120756.5.



Below is a distribution of each of the 5000 runs with respect to the final portfolio value after each run



Graph of profit/loss for each simulation run



The 5% value at risk for this portfolio is 30769.97.

Portfolio 2 Summary and Findings This portfolio consists of 5 equal-weighted ETFS as listed below:

SOXL: Direxion Daily Semiconductor Bull 3x Shares

ROM: ProShares Utlra Technology

DPST: Direxion Daily Regional Banks Bull 3x Shares

DRN: Direxion Daily Real Estate Bull 3x Shares

UCO: ProShares Ultra Bloomberg Crude Oil

After analyzing our portfolio, we found the 5% VAR to be $30.769.97. This value quantifies the extent of possible financial losses for our portfolio over the 20 trading day period simulation. For this particular instance of the simulation, the ending value of our portfolio was 101002.4 dollars.

We have decided to use 5 very aggressive ETFs for this problem, so it makes more sense as to why the VAR for this portfolio is much higher than the VAR for portfolio 1, which consists of safer and more stable ETFs.

Portfolio 3

We have decided to use 5 very aggressive ETFs for this problem, they are listed and imported below: For this portfolio, we have decided to use primarily global real estate and global large cap ETFs.

VNQI: Vanguard Global ex-US Real Estate ETF

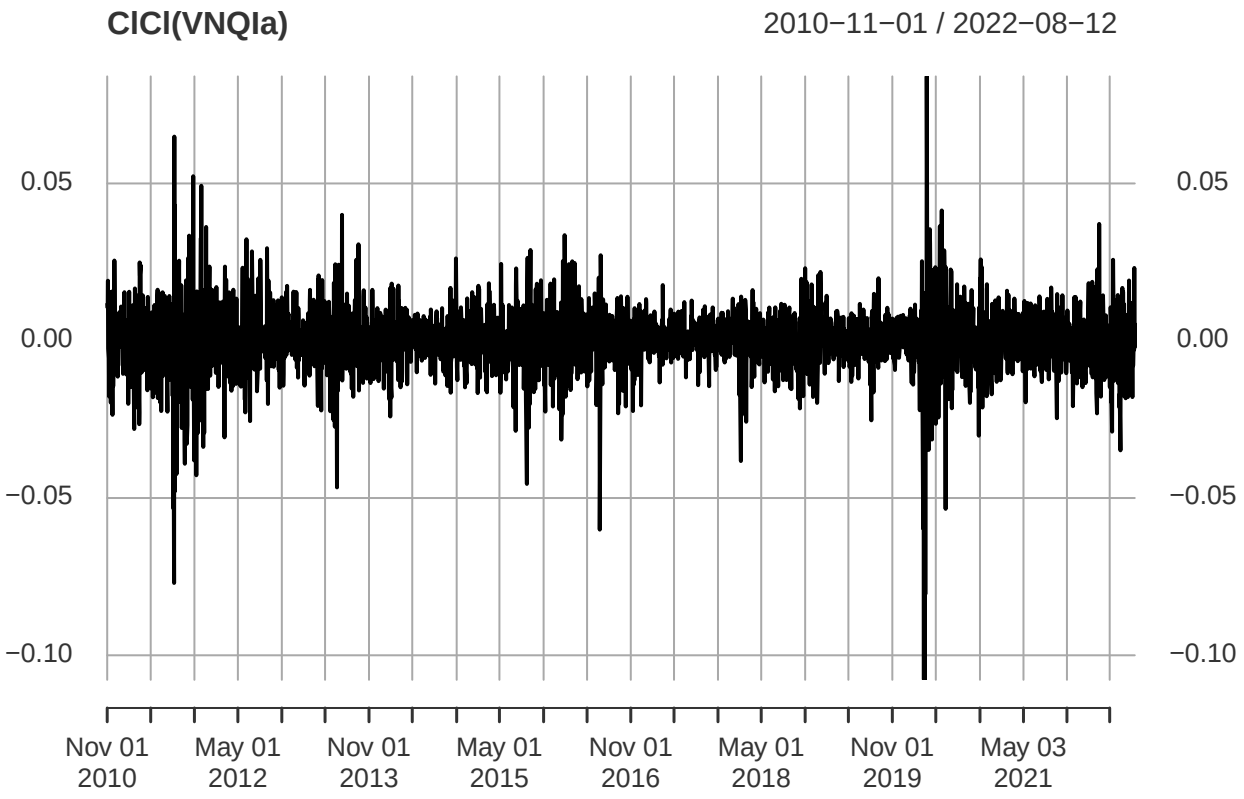
EWJ: iShares MSCI Japan ETF

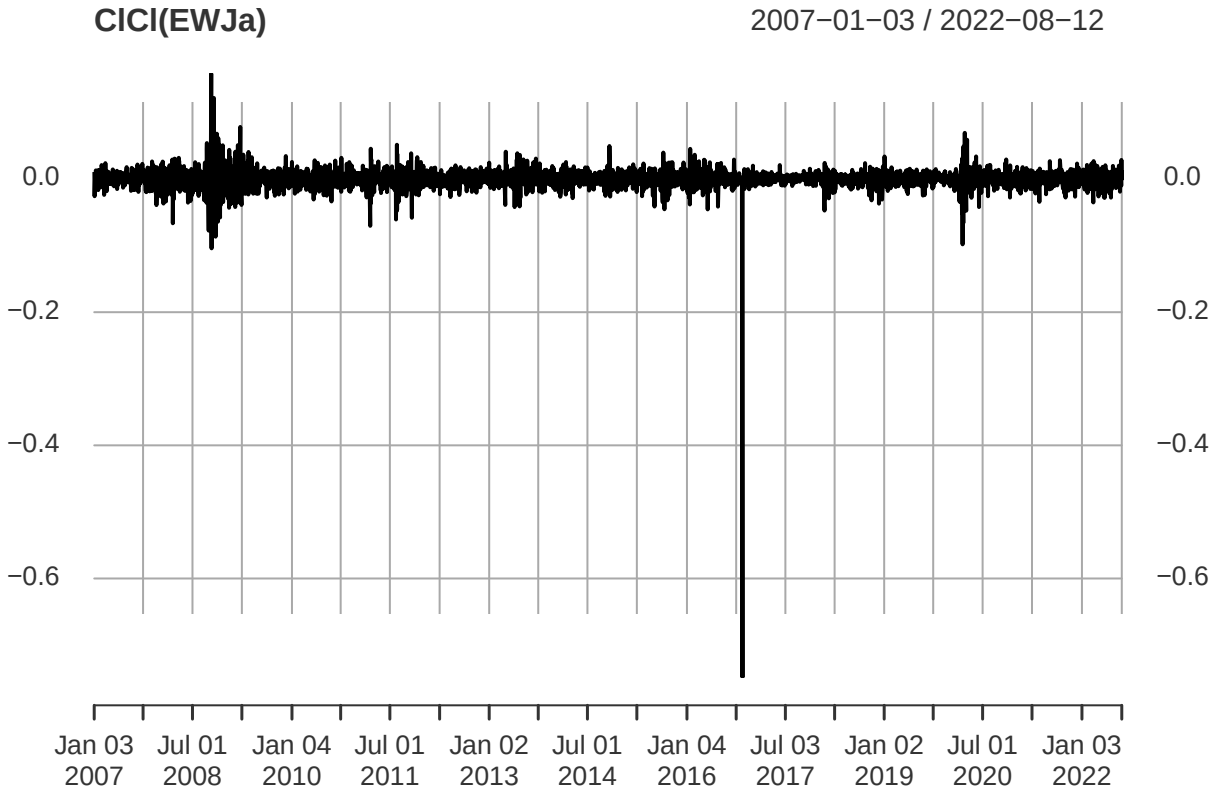
SCHF: Schwab International Equity ETF

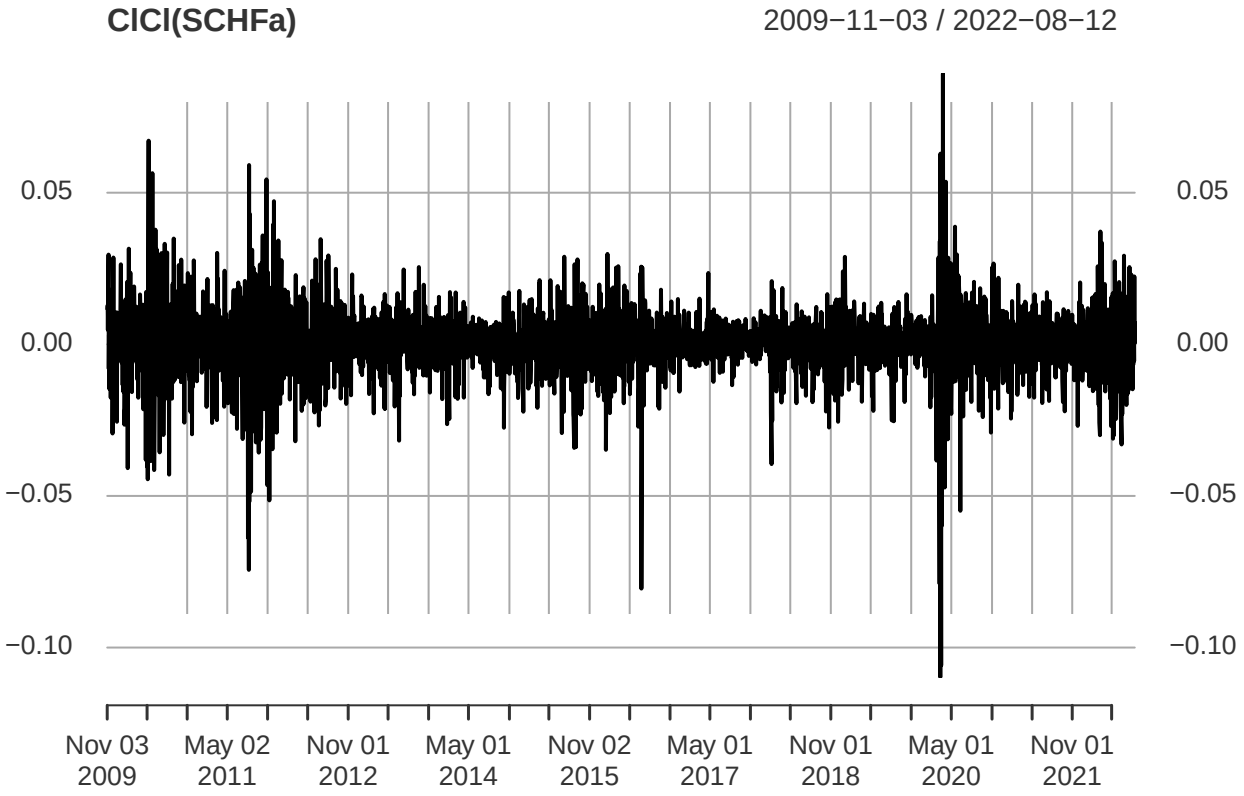
RJN: Elements Rogers International Commodity Index-Energy Total Return ETN

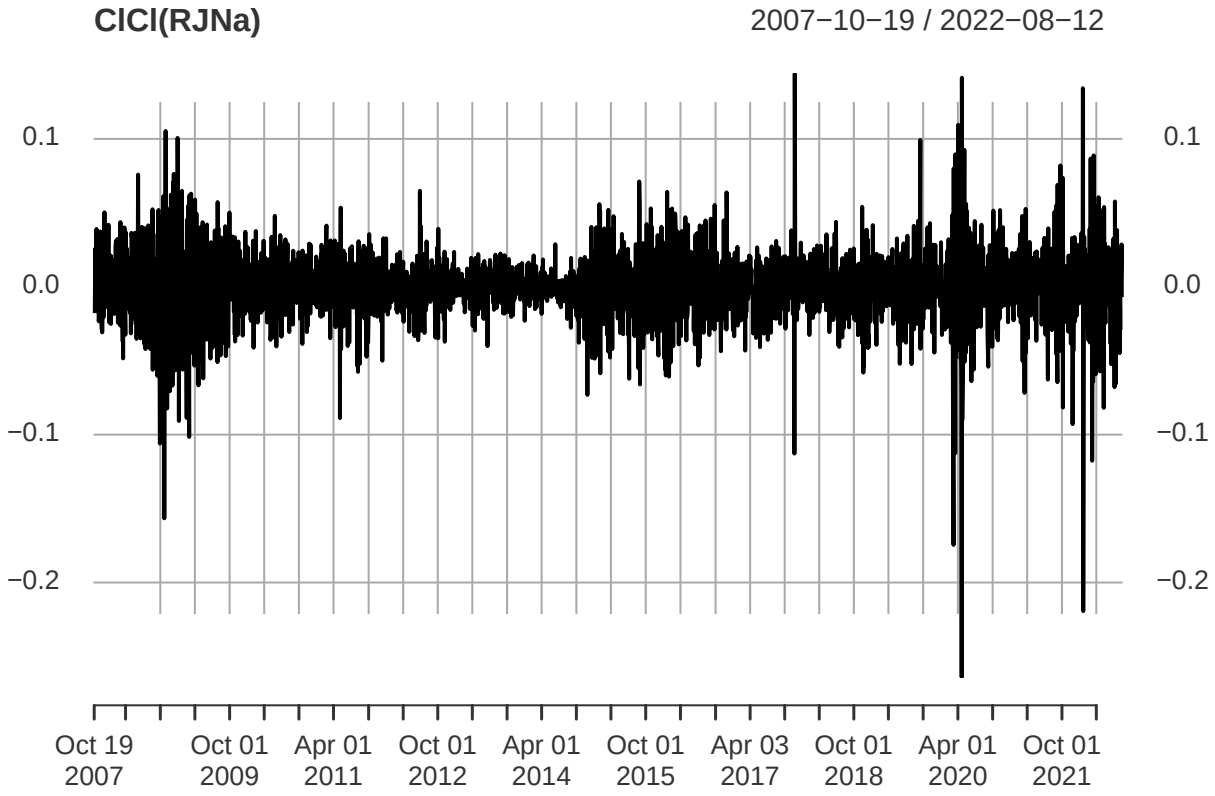
HDEF: Xtrackers MSCI EAFE High Dividend Yield Equity ETF

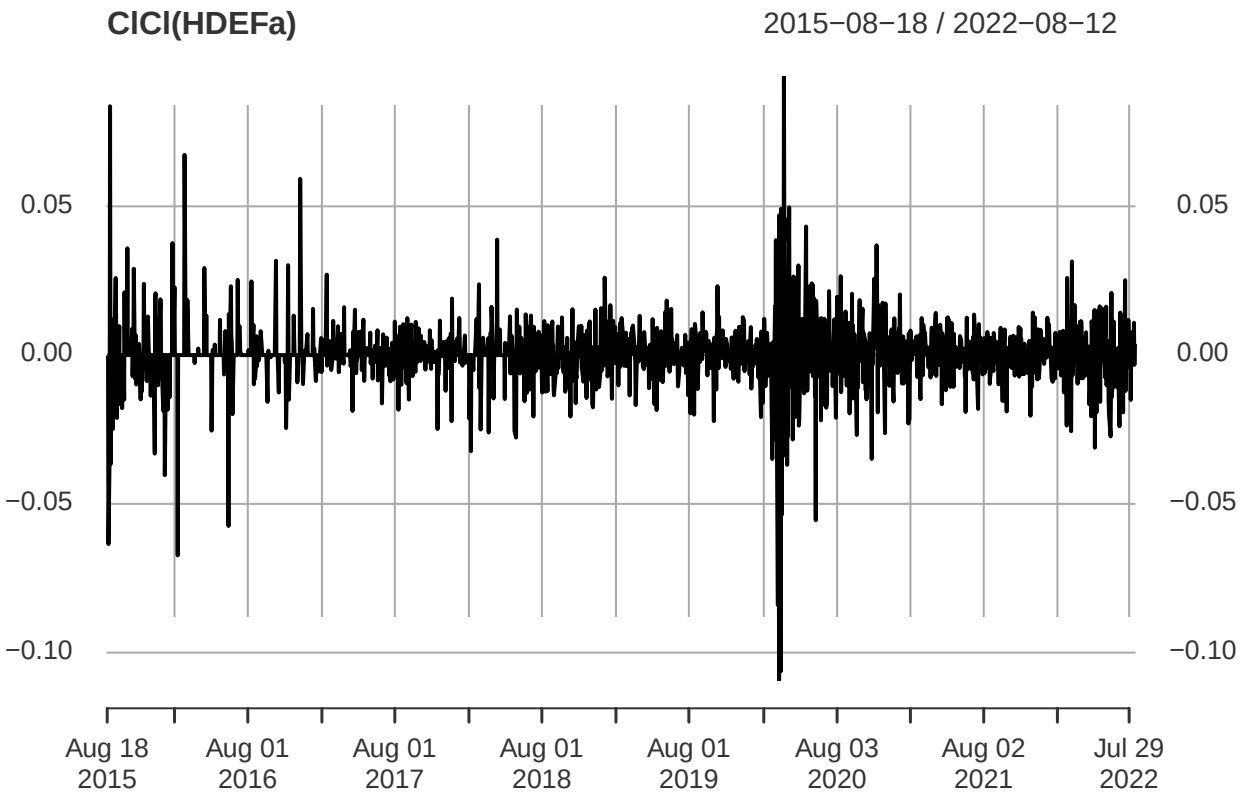
Adjust for splits and dividends Now that the data has been adjusted for splits and dividends, the following graphs display the close-to-close changes for each of the ETFs in our portfolio.



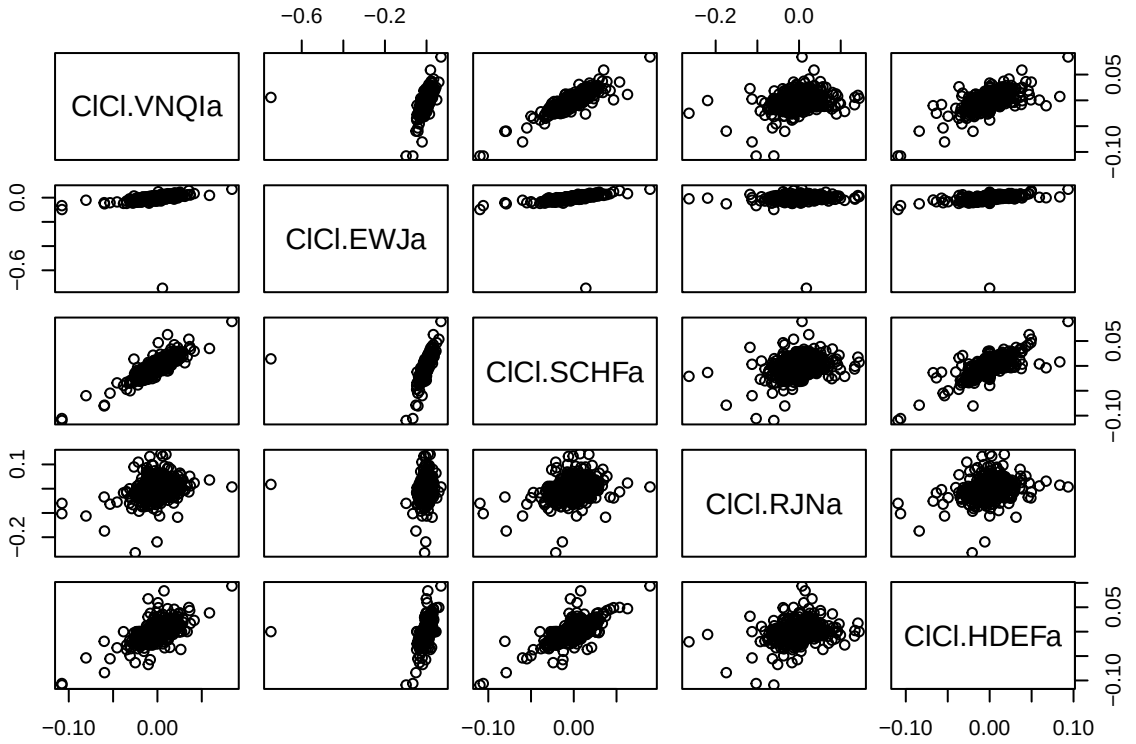

ClCl(EWJa)
2007−01−03 / 2022−08−12
0.0
−0.2
−0.4
−0.6
Jan 03 2007
Jul 01 2008
Jan 04 2010
Jul 01 2011
Jan 02 2013
Jul 01 2014
Jan 04 2016
Jul 03 2017
Jan 02 2019
Jul 01 2020
Jan 03 2022


ClCl(SCHFa)
2009−11−03 / 2022−08−12
0.05
0.00
−0.05
−0.10
Nov 03 2009
May 02 2011
Nov 01 2012
May 01 2014
Nov 02 2015
May 01 2017
Nov 01 2018
May 01 2020
Nov 01 2021


ClCl(RJNa)
2007−10−19 / 2022−08−12
0.1
0.0
−0.1
−0.2
Oct 19
2007
Oct 01
2009
Apr 01
2011
Oct 01
2012
Apr 01
2014
Oct 01
2015
Apr 03
2017
Oct 01
2018
Apr 01
2020
Oct 01
2021

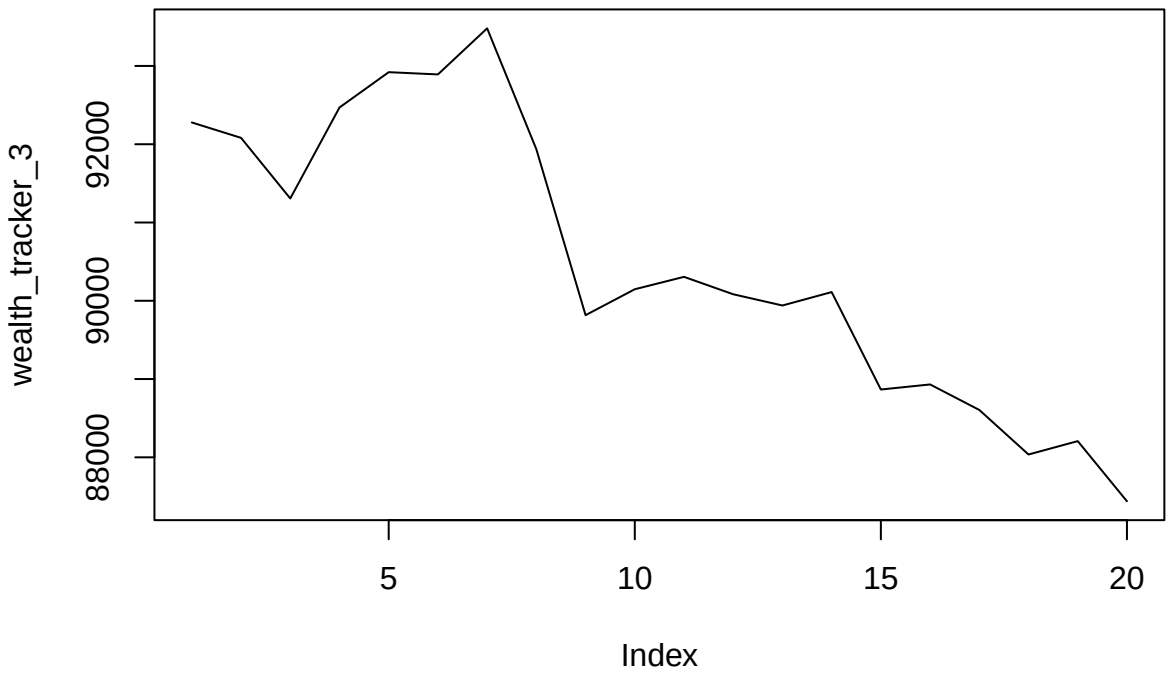


Combine close to close changes in a single matrix Omit all NA in each set so all ETFS start at the same date Once we have found a common start date for the portfolio and got rid of NA values, we are now able to use the pairs function to show the correlation between each of the ETFs in our portfolio.

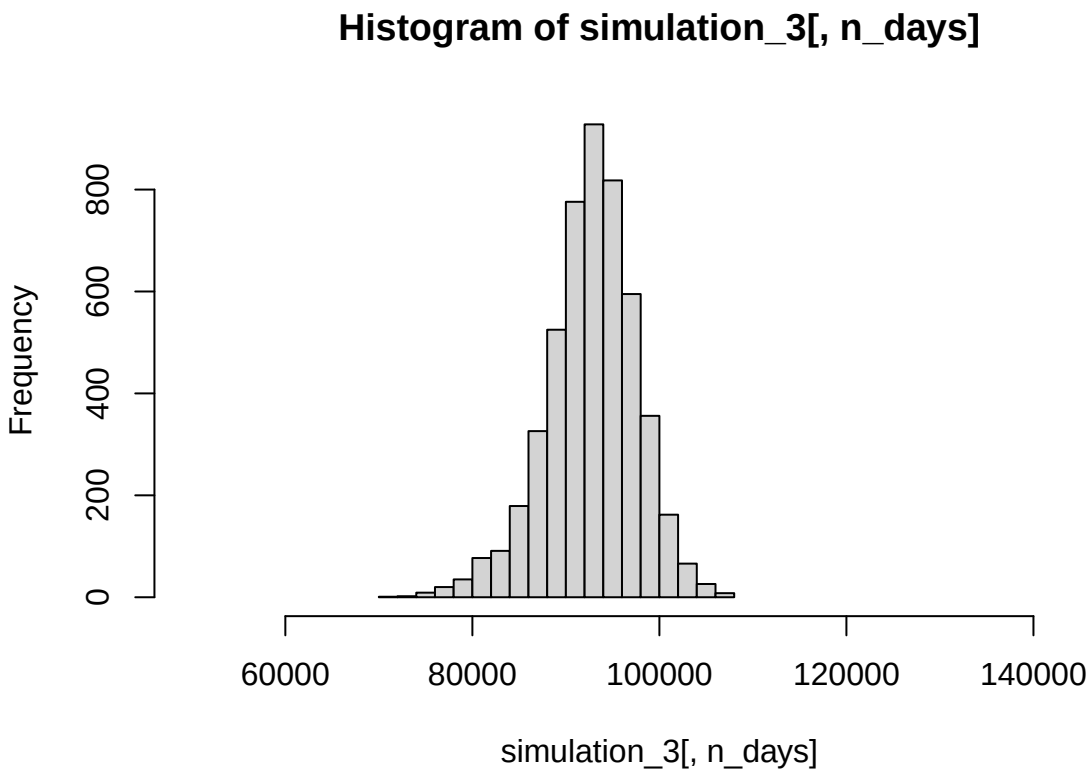


Begin bootstrap approach:

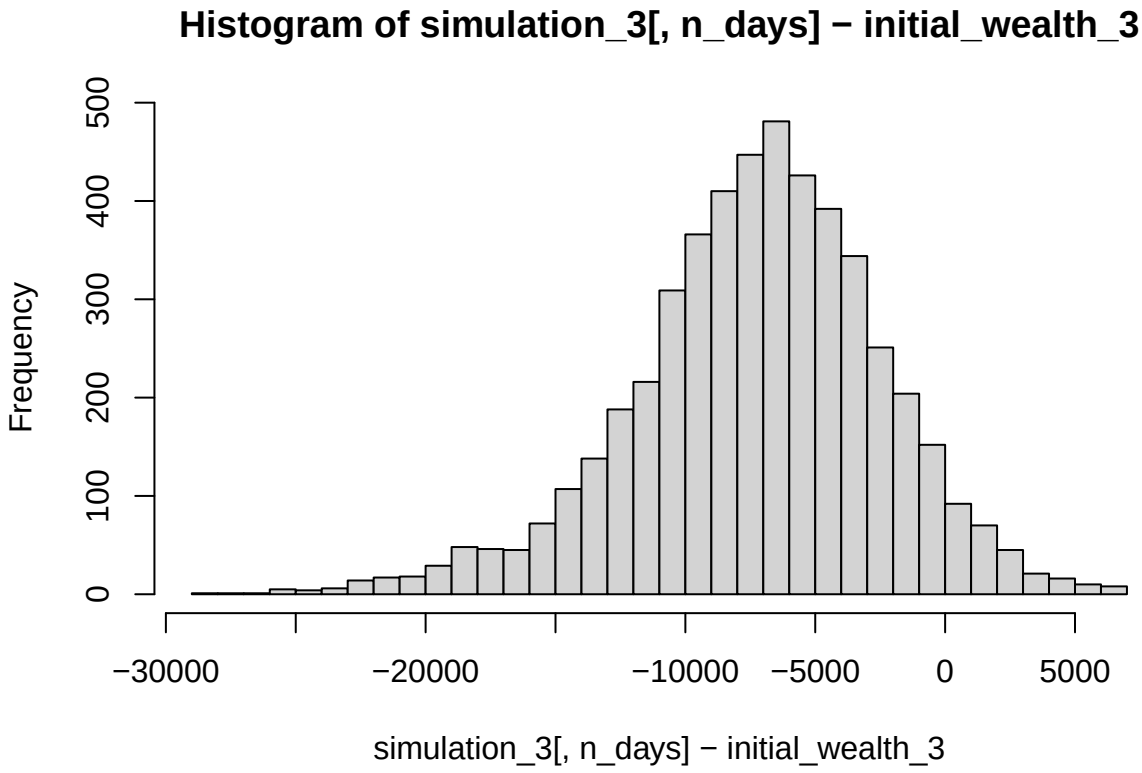
Sample a random return from the empirical joint distribution This simulates a random day Update the value of the holdings to be equal weighted Since we chose 5 ETFS each one will have a weight of 20% Compute our new total wealth Now we will loop over 4 trading weeks, or 20 trading days using the same logic as mapped out above. The graph below shows the average of the 5000 trading simulations over the 20 day time frame as indicated by the question. We can see that this portfolio had an average value of 88628.86.



Below is a distribution of each of the 5000 runs with respect to the final portfolio value after each run



Graph of profit/loss for each simulation run



The 5% value at risk for this portfolio is 16086.42

Portfolio 3 Summary and Findings This portfolio consists of 5 equal-weighted ETFS as listed below:

VNQI: Vanguard Global ex-US Real Estate ETF

EWJ: iShares MSCI Japan ETF

SCHF: Schwab International Equity ETF

RJN: Elements Rogers International Commodity Index-Energy Total Return ETN

HDEF: Xtrackers MSCI EAFE High Dividend Yield Equity ETF

After analyzing our portfolio, we found the 5% VAR to be $16,086.42. This value quantifies the extent of possible financial losses for our portfolio over the 20 trading day period simulation. For this particular instance of the simulation, the ending value of our portfolio was 91925.6 dollars.

For this portfolio, we have decided to use primarily global real estate and global large cap ETFs to simulate other market conditions and economic factors spanning outside the United States.

Question 6: Clustering and PCA

After centering and scaling the data and creating the matrix with Euclidean distances. Different hierarchical methods were tested.

Hierarchical Clustering

Minimum

```
##
## Call:
## hclust(d = distMatrix_wines, method = "single")
##
## Cluster method   : single
## Distance         : euclidean
## Number of objects: 6497
```

Maximum

```
##
## Call:
## hclust(d = distMatrix_wines, method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 6497
```

Average

```
##
## Call:
## hclust(d = distMatrix_wines, method = "average")
##
## Cluster method   : average
## Distance         : euclidean
## Number of objects: 6497
```

Centroid

```
##
## Call:
## hclust(d = distMatrix_wines, method = "centroid")
##
## Cluster method   : centroid
## Distance         : euclidean
## Number of objects: 6497
```

Cutting the trees into clusters

Minimum

```
##    1    2    3
## 6495    1    1
```

This method was not useful because one cluster has 6,493 points and the others have 1 point.

Maximum

```
##    1    2    3
## 6472   24    1
```

Better than using the Minimum, but not good enough because one cluster has 6,469 points and the others have less than 25.

Average

```
##    1    2    3    4    5
## 6464    6   25    1    1
```

Better than using the Minimum or Maximum, but not good enough because one cluster has 6,464 points and the others have 25 or less.

Centroid

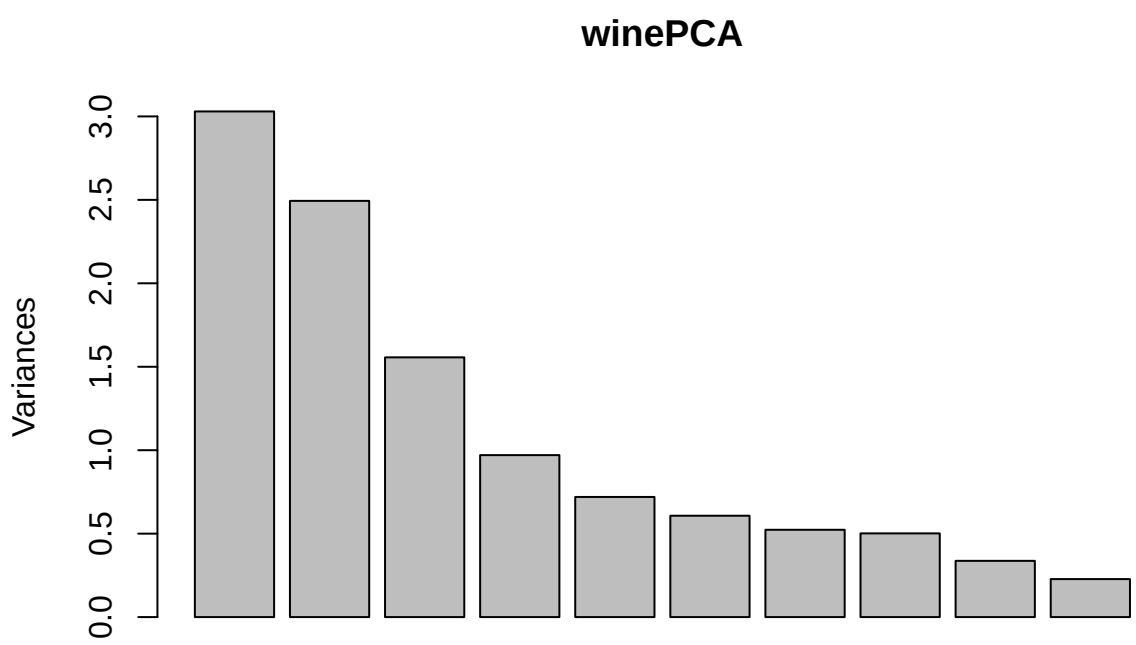
```
##    1    2    3
## 6495    1    1
```

```
##    1    2    3    4    5
## 6490    4    1    1    1
```

Not good enough because three of the clusters have one point each.

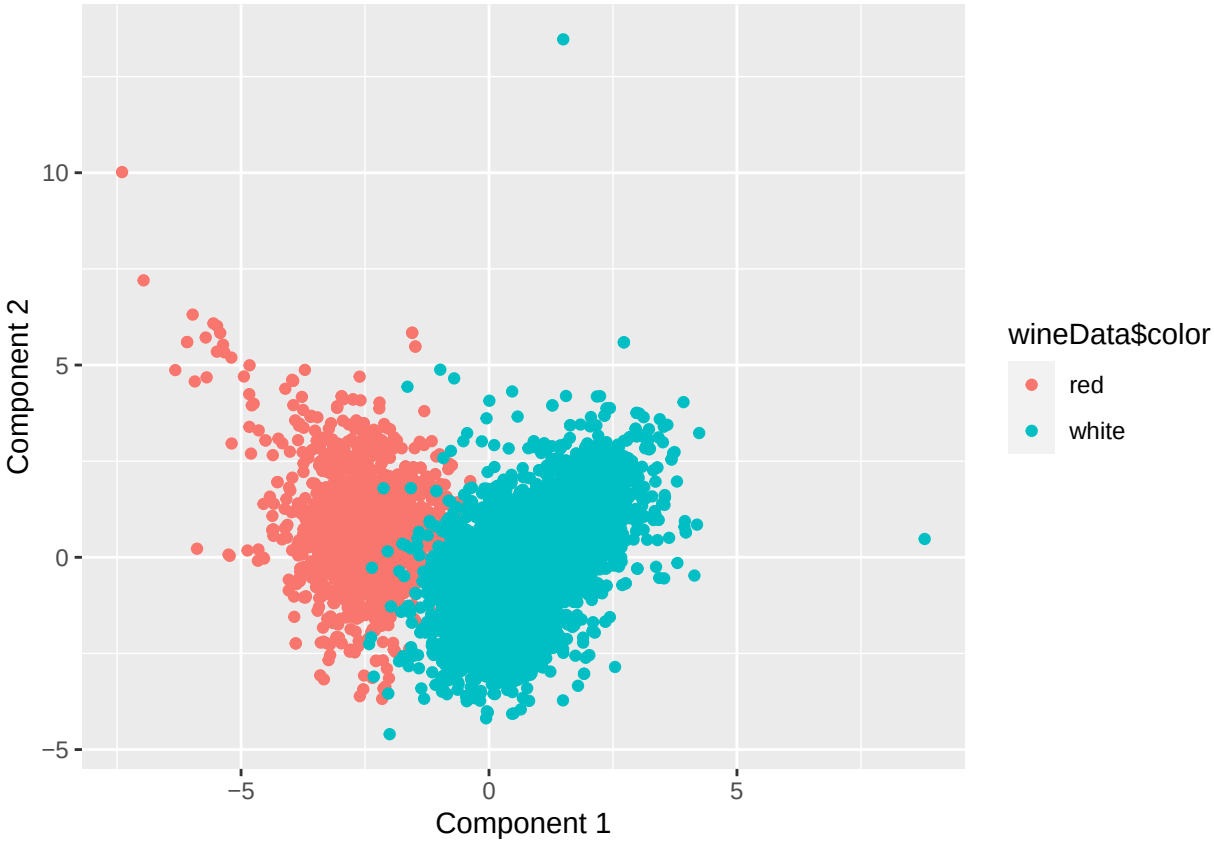
PCA

After centering and scaling the data and setting up the PCA of the properties of the wines, a variance plot was created.

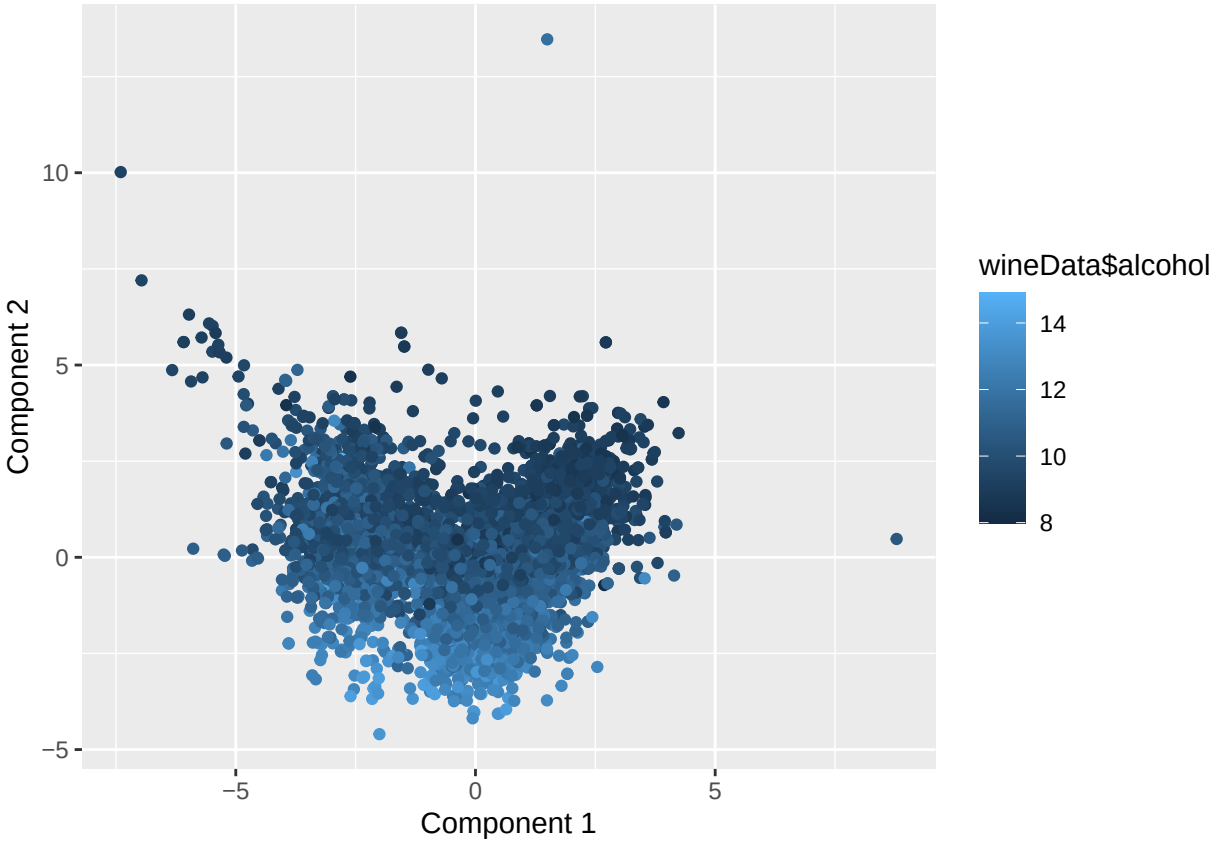



```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.90009
##                            PC8     PC9   PC10    PC11
## Standard deviation     0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.94568 0.97632 0.9970 1.00000
```

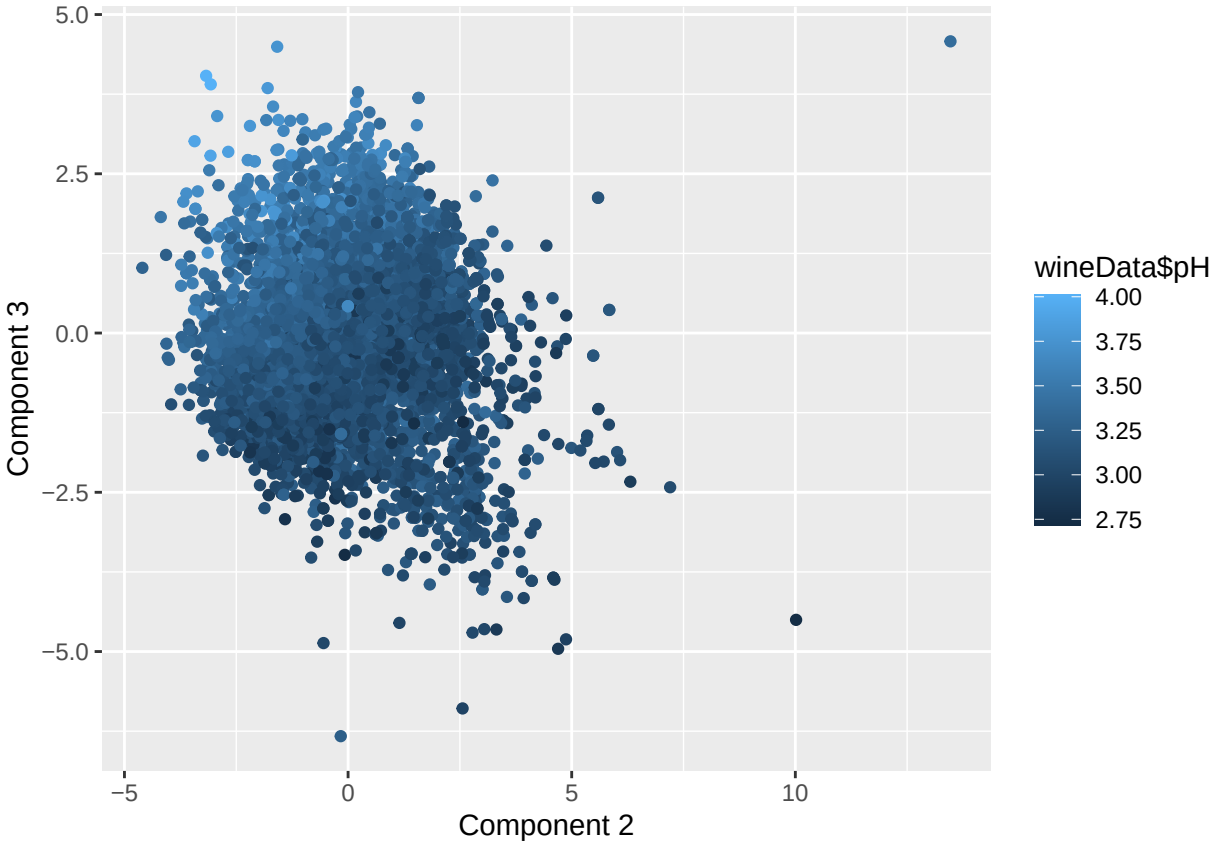
Cumulative Proportion: The first 5 models account for 79.73% of the variation.



It can be seen that the principal component 1 is dividing wines into red and white.



It can be seen that the principal component 2 is dividing wines by alcohol level.



It can be seen that the principal component 3 is dividing wines by pH.

Ordering Loadings

PC1 Getting the top 2 most positive chemical properties in PC1. This are the White Wine chemical properties:

```
## [1] "total.sulfur.dioxide" "free.sulfur.dioxide"
```

Getting the top 2 most negative chemical properties in PC1. This are the Red Wine chemical properties:

PC2 Getting the top 2 most positive chemical properties in PC2. This are the chemical properties of the lower alcohol levels:

```
## [1] "density"       "fixed.acidity"
```

Getting the top 2 most negative chemical properties in PC2. This are the chemical properties of the higher alcohol levels:

PC3 Getting the top 2 most positive chemical properties in PC2. This are the chemical properties of high pH levels:

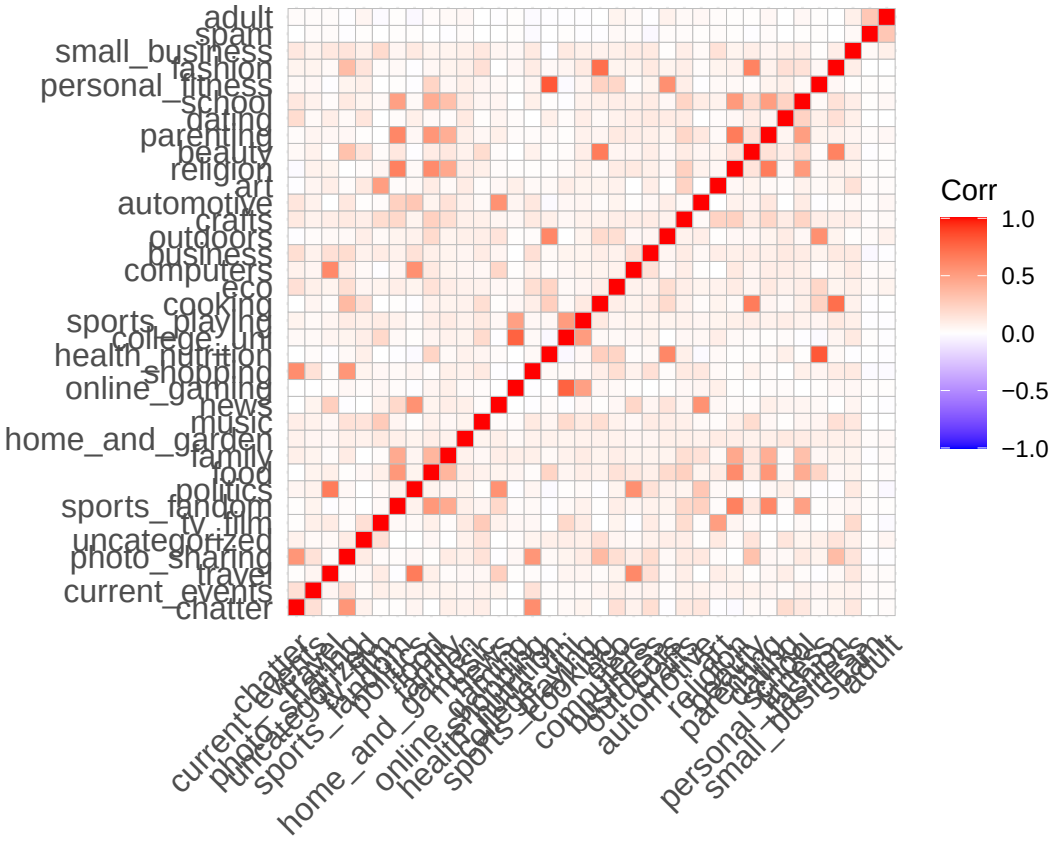
```
## [1] "pH"               "volatile.acidity"
```

Getting the top 2 most negative chemical properties in PC2. This are the chemical properties of low pH levels:

It can be seen that in this case, the PCA algorithm works better for a dimensionality reduction. With PC1 it can be differentiated between red and white wines.

Question 7: Market Segmentation

Below is the initial correlation between all categories before we alter and clean the data set.



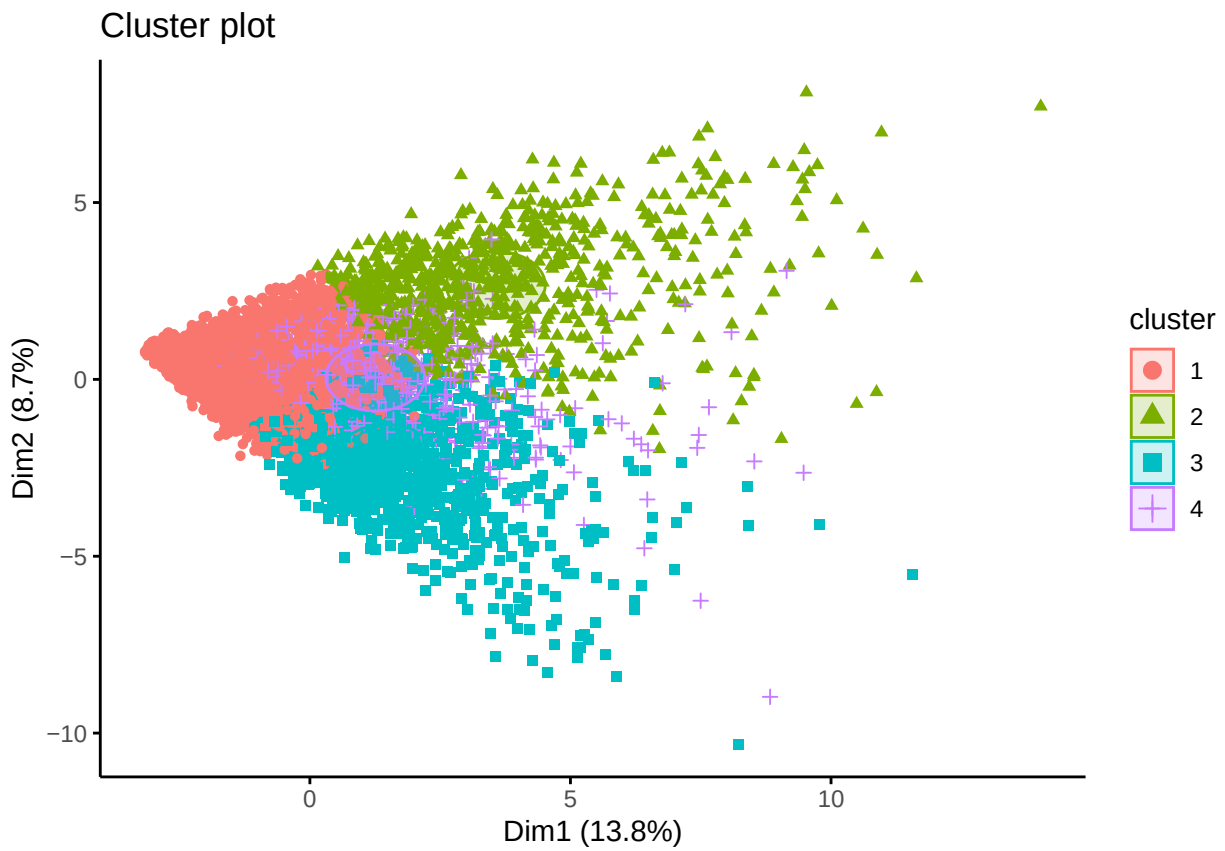
Trends observed from the correlation plot: personal_fitness and health_nutrition seem to have a strong positive correlation beauty and fashion are seen to have a positive correlation travel and small_business have a slight positive correlation

Given the description of the plethora of twitter bots on the platform, we decided to remove the “spam” and “adult” columns from our data set Furthermore, we reviewed the “uncategorized” label as well as the “chatter” categories from our set due to the fact that we fill these unspecified tweets will provide minimal context and accuracy to our analysis Once we have cleaned and accounted for all columns in which we do not want for our analysis, we then cleaned and scaled the data:

Now we will extract the centers and scales from the newly rescaled data set

We will perform k-means clustering on our data set to see the different clusters of market segmentation in our data set. The elbow curve is displayed below to determine the optimal amount of clusters

We have decided to proceed with 4 clusters and use that to see if we can find any info about the market segementation per tweet.



From this final table of t_clust_results we are able to now observe a distinct set of words/categories for each cluster based on the mean of other categories and how they vary across different clusters. If we look in to the t_clust_results table, we are able to see that each of the 4 clusters has personalized attributes affiliated with each of them.

For example in cluster 3, some of the highest categories include 'outdoors','personal fitness', and 'health nutrition'. These categories make it clear that most of the tweets in cluster 3 have to do with fitness, exercising, and enjoying the outdoors.

Furthermore in cluster 2, some of the highest categories include 'school','parenting','religion', and 'food'. These categories indicate that the tweets in cluster 2 have to do more with food, culture, and education, all topics that are important to most families.

Question 8: The Reuters Corpus

Question

Similar to our example in class, we hope to dissect an author's documents and try our best to group them by their contents. More specifically, we'd like to answer the question of what terms appear most frequently in author Heather Scoffield's documents, as well as how related these documents are in space.

Approach

To answer this question, we will start by reading each of Heather Scoffield's documents into a file list. To make interpretation and recognition easier, we can clean up the original document names and applied them to the file list. We then plan to follow general tokenization steps, such as removing numbers, punctuation,

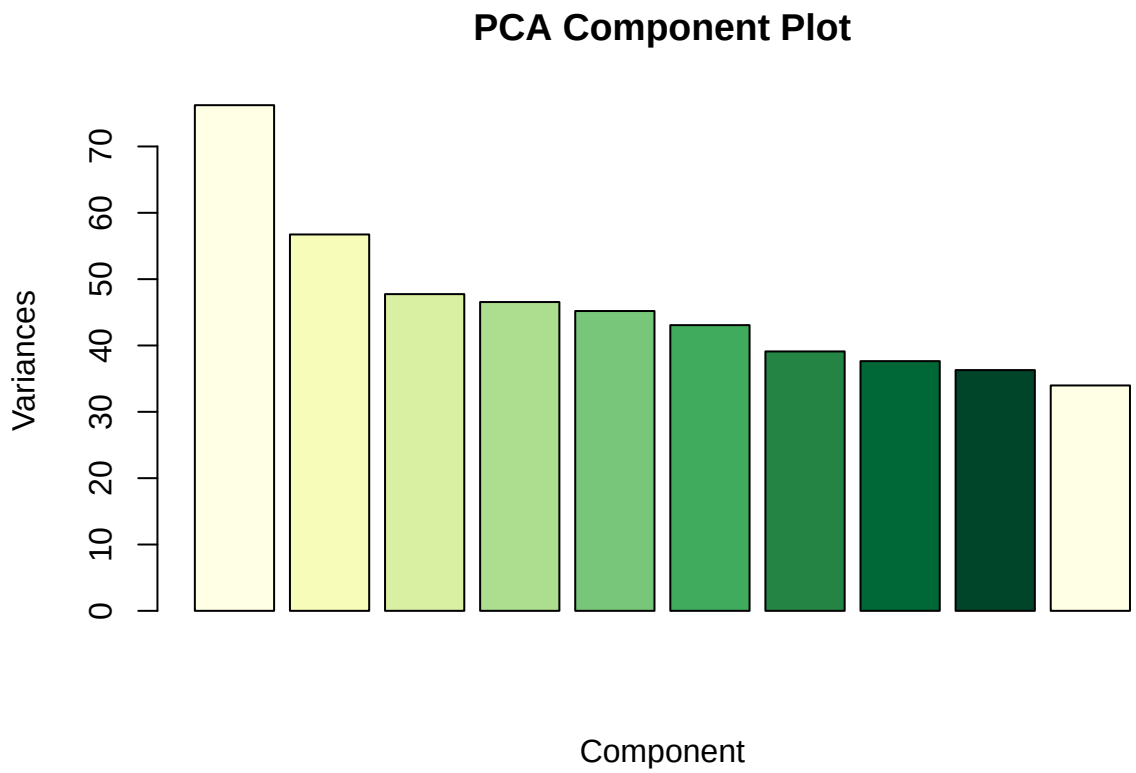
white spaces, and common stop words. This whole process ensures we are analyzing relevant data and removes unnecessary words from each document.

The next step is to create a document-term matrix (DTM), which contains the frequency of terms in each of the documents. The corresponding inverse document frequency (IDF) assigns higher weights to uncommon terms across the corpus as a whole. Finally, a term frequency-inverse document frequency (TF-IDF) will be obtained by combining the DTM and IDF.

Finally, we will perform principal component analysis (PCA) to see if we can summarize each document within the corpus given its qualities. Additionally, we will perform clustering of the documents to see which ones are most similar.

Results

After completing PCA, we have summarized each document in the corpus into a single pair of values in the real coordinate space. In the plot below, we can see that we were able to maintain a majority of the variation between documents within the first 2 principal components. Words that appeared 75 or more times within the DTM are listed below. Furthermore, examining the loadings will give us an idea of which words have the strongest influence on each component.




```
##  [1] "character"   "companies"   "company"     "exploration" "million"
##  [6] "mining"      "percent"     "said"        "will"        "year"
## [11] "analysts"    "corp"        "gold"        "government"  "stock"
## [16] "barrick"     "brex"        "busang"      "indonesian"
```

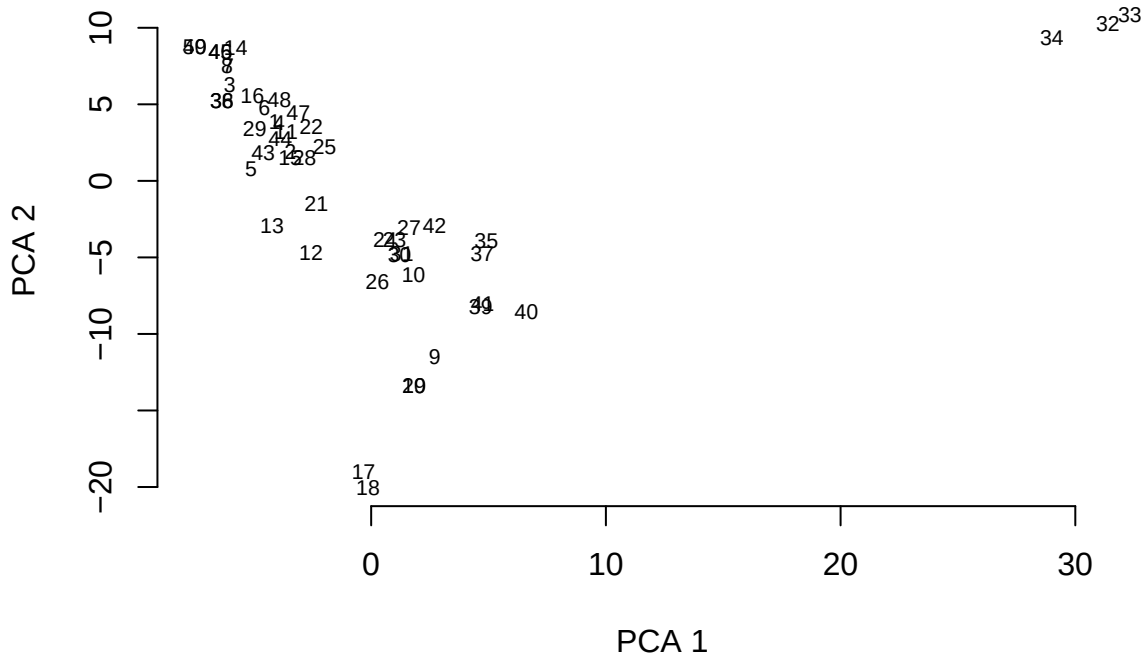
```
##    deadline      blocks     busangs      clinch     clocked    crossing
```

```
##   0.1037899   0.1034938   0.1034938   0.1034938   0.1034938   0.1034938
##    directed discouraged     dotting    expedite
##   0.1034938   0.1034938   0.1034938   0.1034938

##        brex     minorca  indonesian    suhartos   contracts       horst
## -0.09406477 -0.08760043 -0.08734735 -0.08572991 -0.08527286 -0.08497220
##      roland   indonesia    partners         son
## -0.08460859 -0.08315992 -0.08307677 -0.08111043
```

Now we can examine the magnitude of the first 2 principal components for each document. Furthermore, we can see how the documents were placed in the real coordinate space and try to see what qualities they have. From the plot below, we can see that documents 32, 33, and 34 share similar characteristics, as well as documents 17 and 18. Let's try and get a summary of their content.

```
##
## Docs       PC1         PC2
##   1  -4.111420   3.8544729
##   2  -3.438514   1.9221374
##   3  -6.019950   6.3068271
##   4  -3.917225   3.7813299
##   5  -5.122433   0.7673308
##   6  -4.549808   4.7863933
##   7  -6.102428   7.5450820
##   8  -6.136413   7.5013545
##   9   2.715296 -11.4886497
##   10  1.808049  -6.1213339
```



```
#view first group of docs
content(heather[[32]])[1:3]
```

```
## [1] "No final deal was in sight Wednesday for Bre-X Minerals Ltd. and Barrick Gold Corp., which are
## [2] "As a Wednesday deadline slid by, Bre-X and Barrick said they were still trying to hammer out se
## [3] "\"Several points remain outstanding,\" said Barrick spokesman Vince Borg. \"An overall deal has
```

```
content(heather[[33]])[1:3]
```

```
## [1] "Investors were on edge on Wednesday, anxiously awaiting the outcome of talks between Canada's B
## [2] "As a Dec. 4 deadline slid by, Bre-X and Barrick said they were still trying to work out several
## [3] "\"Several points remain outstanding,\" Barrick spokesman Vince Borg said. \"An overall deal has
```

```
content(heather[[34]])[1:3]
```

```
## [1] "No final deal was in sight Wednesday for Bre-X Minerals Ltd. and Barrick Gold Corp., which are
## [2] "As a Wednesday deadline slid by, Bre-X and Barrick said they were still trying to hammer out se
## [3] "A few issues remain to be solved, and Bre-X will have more news on the negotiations \"shortly,\
```

Documents 32, 33, and 34 discuss a business deal between Bre-X Minerals Ltd, Barrick Gold Corp., and Indonesia's Busang gold deposits. These documents appear to be recording relevant updates between mining companies and the Indonesian government.

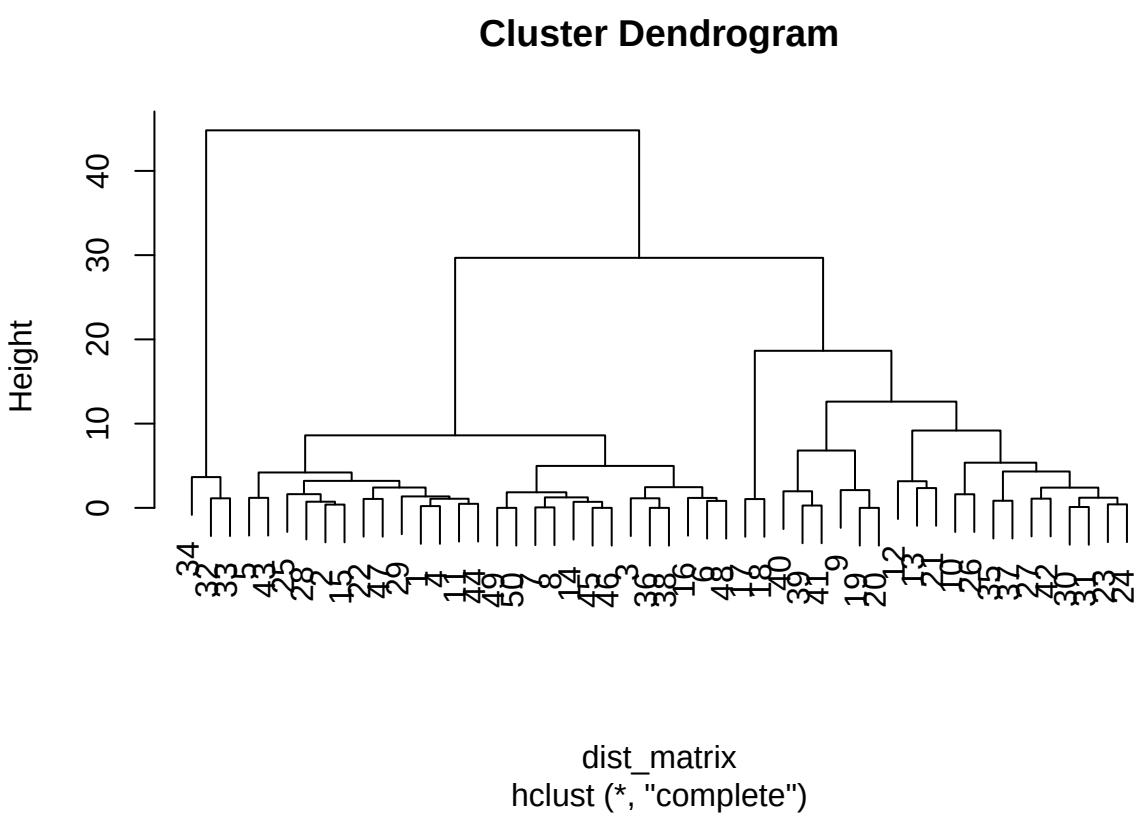
```
#view second group of docs
content(heather[[17]])[1:3]
```

```
## [1] "Canadian mining company Bre-X Minerals Ltd has been in hiding since it announced 17 days ago tha
## [2] "Despite a constant whirl of rumors and persistent questions that have sent the company's shares
## [3] "The Calgary-based company that controls one of the world's biggest gold prospects in Indonesia
```

```
content(heather[[18]])[1:3]
```

```
## [1] "Bre-X Minerals Ltd. has been silent since it said last month that it formed a partnership with
## [2] "Despite a whirl of rumours and persistent questions that have sent the Canadian mining company'
## [3] "The Calgary-based company that controls one of the world's biggest gold prospects in Indonesia
```

Alternatively, documents 17 and 18 focus more on Bre-X Minerals Ltd and their financials. We can now perform clustering using the PCA scores obtained above to see if we can uncover additional relationships between the documents.




```
#looking at the 5th cluster, we see that it placed documents similarly
which(pruned_heather == 5)
```

```
## 32 33 34
## 32 33 34
```

```
#same with 4th cluster
which(pruned_heather == 4)
```

```
## 17 18
## 17 18
```

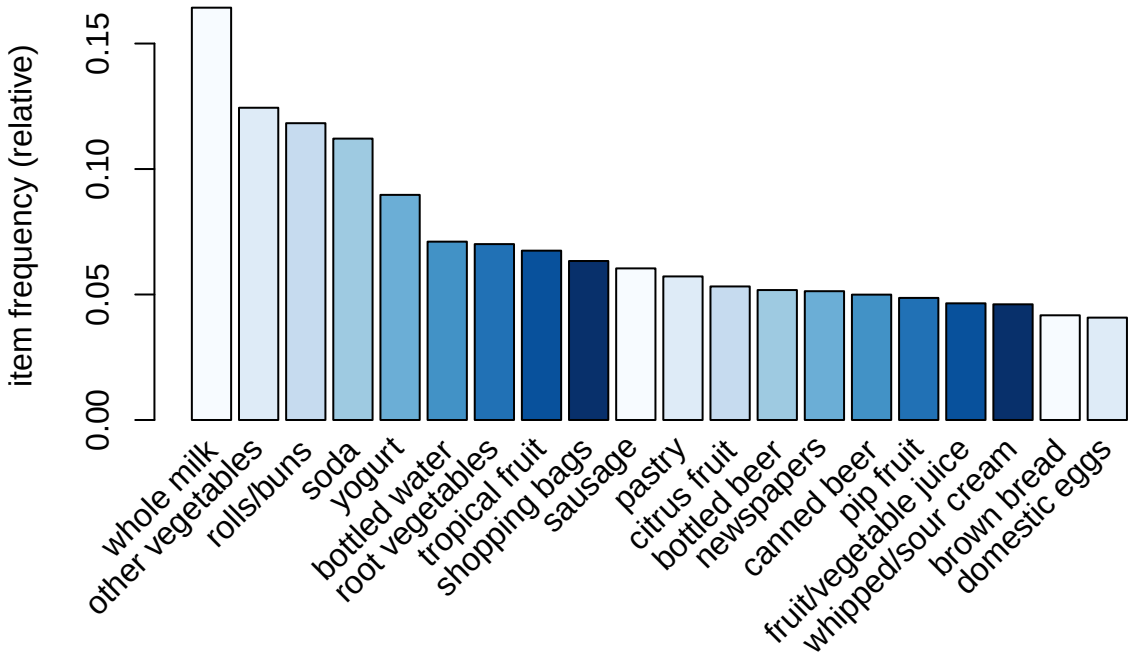
Clustering has resulted in similar groupings to what we saw in PCA. Documents 32, 33, and 34 are grouped together, and so are documents 17 and 18 (as expected). Although the results appear to be essentially the same as before, examining a dendrogram can be a simpler way to see how the documents are related.

Conclusion

In conclusion, it's clear that PCA and clustering are optimal for organizing the documents within the corpus. PCA allowed us to summarize each document into a single pair of numbers while still maintaining a majority of the variation in the dataset. While it may be difficult to understand the contents of these documents from our perspective, it's likely that this grouping can be advantageous for companies looking to organize articles/documents regarding their performance and business deals. This example demonstrates how we can take the composition of words within a document and use them to make informed decisions.

Question 9: Association Rule Mining

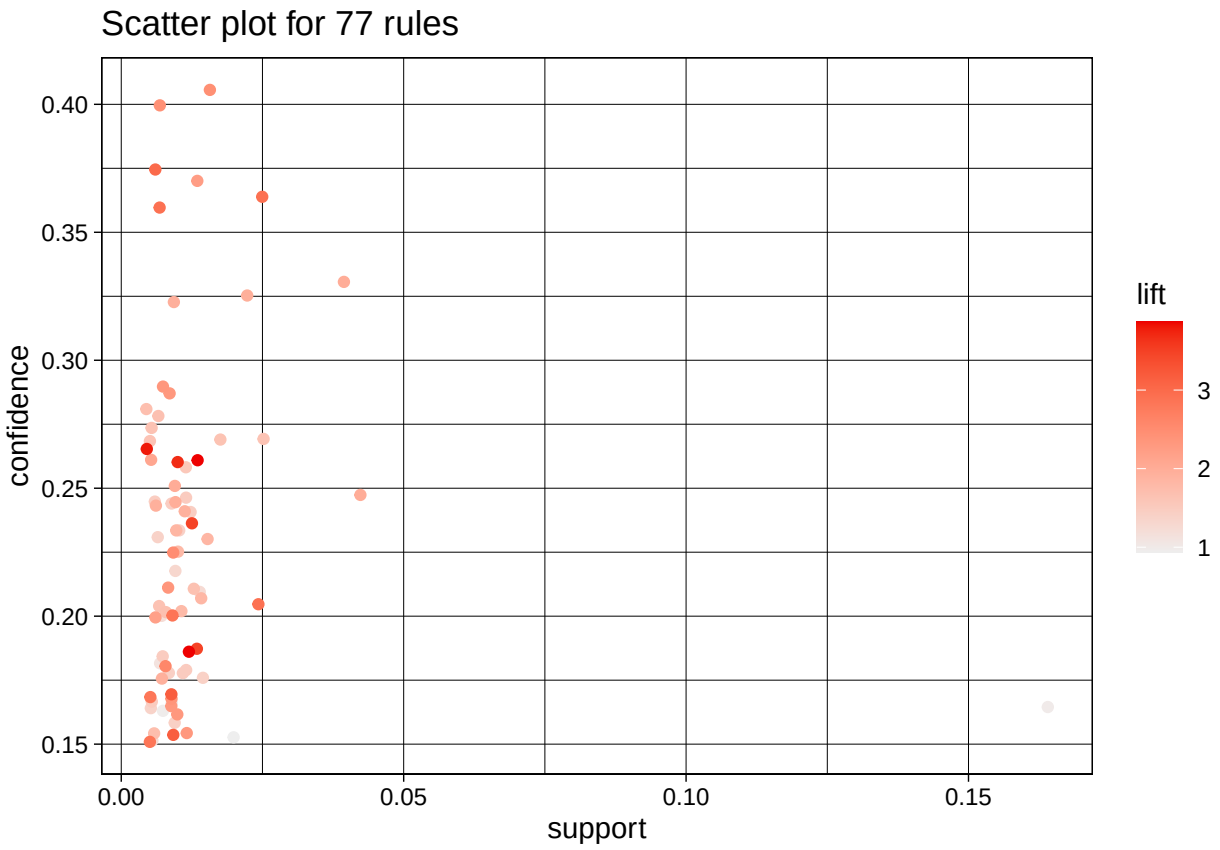
Using data regarding different customer's baskets at the grocery store allows us to predict what will be purchased by future customers. We began by reading in the dataset and performing data wrangling to get it in the correct format. This was done by creating a new ID variable for each customer, then pivoting longer to get each customer's basket items into their own observation. We can now split the data into baskets organized by customer and prepare to feed it into the Apriori algorithm. First, let's take a look at the most frequent cart items within the network.



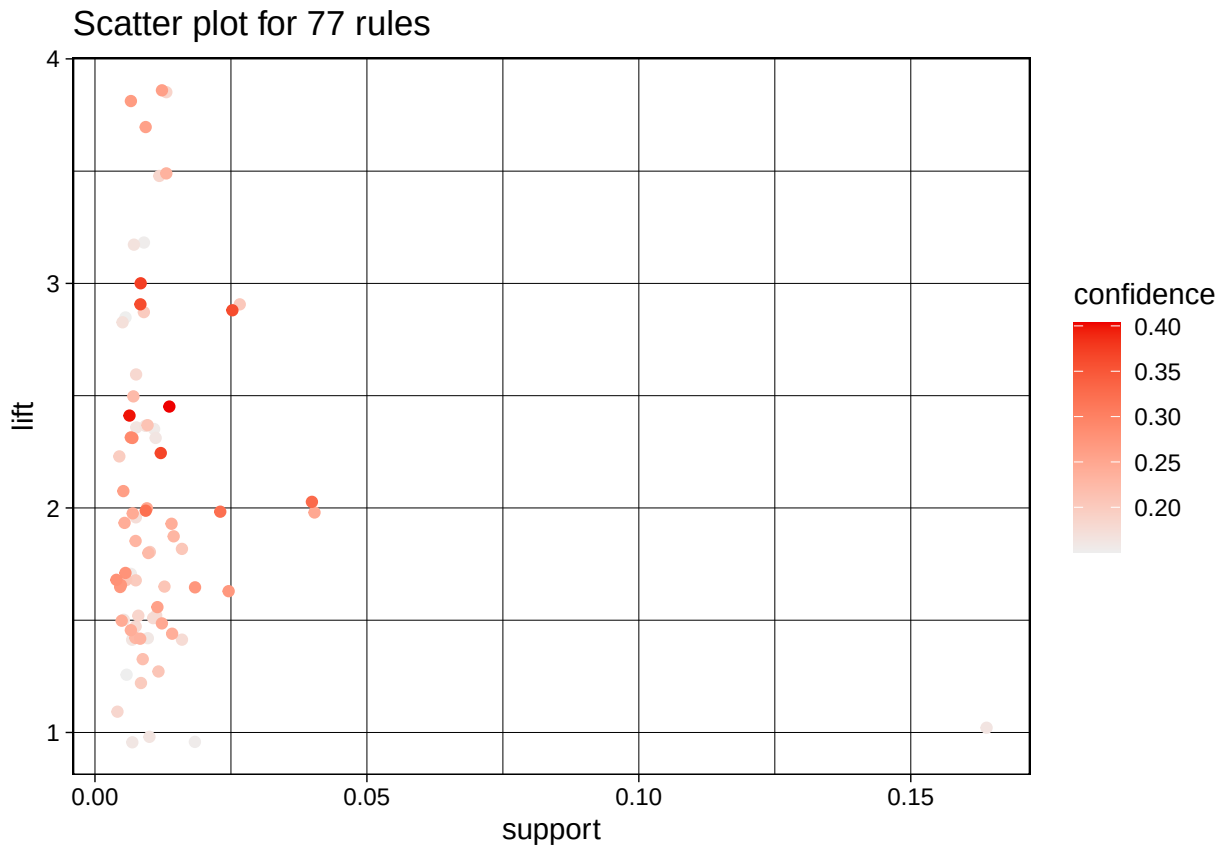
We can see that the item most commonly in our customer's baskets is whole milk, followed by other vegetables, rolls/buns, and soda. Now that we can visualize each item's relative frequency, we can now create our association rules. When determining the parameters for the algorithm, we first started with a low threshold for each rule (support >= 0.005, confidence >= 0.15, maxlen = 5). This is to ensure we are only selecting rules that will give us relative information about the network. Seen below is a scatterplot of each generated rule, which helps us visualize the support, confidence, and lift for each rule.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.15    0.1    1 none FALSE            TRUE       5   0.005      1
##  maxlen target  ext
##       5  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
```

```
##
## Absolute minimum support count: 76
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 15296 transaction(s)] done [0.01s].
## sorting and recoding items ... [101 item(s)] done [0.00s].
## creating transaction tree ... done [0.01s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [77 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```



The plot supports the idea that rules with high lift tend to have low support. We put lift on the y-axis to get a different perspective:

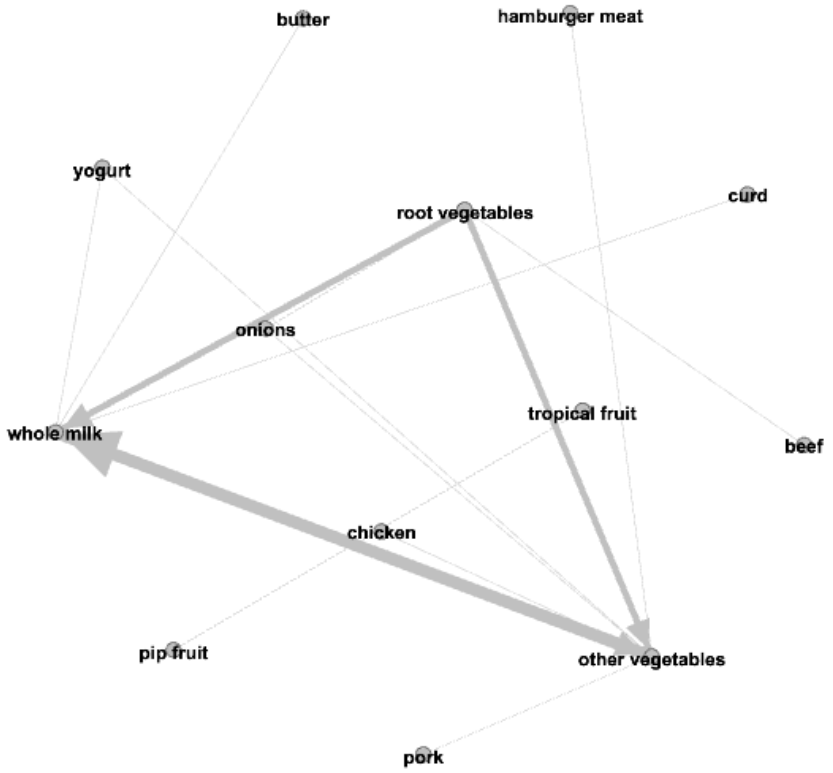


This plot helps us choose stricter parameters for the graph we will export to Gephi. We chose a subset of rules with lift greater than 1.75 and confidence greater than 0.25, which filters for the most important rules in the network. The resulting 16 rules can be seen below.

```
##      lhs                                    rhs                support
## [1]  {onions}                            => {root vegetables}  0.005295502
## [2]  {onions}                            => {other vegetables} 0.007452929
## [3]  {hamburger meat}                    => {other vegetables} 0.006210774
## [4]  {chicken}                           => {other vegetables} 0.007975941
## [5]  {beef}                              => {root vegetables}  0.008695084
## [6]  {curd}                              => {whole milk}       0.012617678
## [7]  {butter}                            => {whole milk}       0.014382845
## [8]  {pork}                              => {other vegetables} 0.009283473
## [9]  {pip fruit}                         => {tropical fruit}   0.012683054
## [10] {root vegetables}                   => {other vegetables} 0.025366109
## [11] {root vegetables}                   => {whole milk}       0.022620293
## [12] {other vegetables}                  => {whole milk}       0.040860356
## [13] {other vegetables, root vegetables} => {whole milk}       0.008172071
## [14] {root vegetables, whole milk}       => {other vegetables} 0.008172071
## [15] {other vegetables, yogurt}          => {whole milk}       0.006341527
## [16] {whole milk, yogurt}                => {other vegetables} 0.006341527
##      confidence coverage   lift     count
## [1]  0.2655738  0.01993985 3.789381  81
## [2]  0.3737705  0.01993985 3.004306 114
## [3]  0.2905199  0.02137814 2.335151  95
## [4]  0.2890995  0.02758891 2.323734 122
```

```
## [5]  0.2577519  0.03373431 3.677774 133
## [6]  0.3683206  0.03425732 2.241875 193
## [7]  0.4036697  0.03563023 2.457036 220
## [8]  0.2504409  0.03706851 2.013003 142
## [9]  0.2607527  0.04864017 3.864800 194
## [10] 0.3619403  0.07008368 2.909216 388
## [11] 0.3227612  0.07008368 1.964566 346
## [12] 0.3284288  0.12441161 1.999064 625
## [13] 0.3221649  0.02536611 1.960937 125
## [14] 0.3612717  0.02262029 2.903842 125
## [15] 0.3991770  0.01588651 2.429690  97
## [16] 0.2614555  0.02425471 2.101536  97
```

We can see that our highest lift rule is from onions to root vegetables, which suggests that customers who add onions to their cart are more likely to add root vegetables to their basket. Finally, we export the network as a .graphml file so that it can be viewed and interpreted in Gephi.



The resulting network, although appearing quite simple, explains some relationships between the items in the basket of each customer. We can see that customers who add root/other vegetables to their baskets often add whole milk as well. Furthermore, we can see other cold goods such as yogurt and butter that are connected to whole milk. These relationships are easily interpretable since since our network consists of only a few grocery items. We expect customers who buy whole milk to add other cold items next, which is why we see yogurt and butter connected to milk. Additionally, the larger arrows pointing to whole milk leads us to believe that most people begin their shopping trip at the front of the grocery store where produce is often

located before adding other items to their cart.

With the addition of more customer's baskets and more grocery items, this network could become much more elaborate and offer further interesting relationships regarding customer's grocery baskets. The reason we see the greatest magnitude of vegetables and milk is likely due to the fact that these are some of the most commonly purchased groceries for all customers.